Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in

Scienze Statistiche

Confronto di indicatori dell'impatto scientifico di un autore: previsione del tempo fino al raggiungimento del ruolo di Professore ordinario

Relatore:

Prof. Bruno Scarpa

Co-relatore:

Dott. Tommaso Dorigo

Laureanda:

Elvia Arcolin

Matricola:

2007701

Anno Accademico 2022/2023

Indice

Introduzione

In questa tesi ci si pone l'obiettivo di impiegare gli indicatori bibliometrici come strumento predittivo della durata delle carriere accademiche e confrontare la loro utilità in termini di strumenti di valutazione. Considerando un campione di Professori ordinari al 31/12/2021 e un campione di Ricercatori al 31/12/2005 in Italia, con le loro rispettive produzioni scientifiche e ruoli accademici occupati nell'intero corso di carriera, si vogliono utilizzare alcuni indici bibliometrici per la previsione delle durate di carriera, dalla prima pubblicazione fino al raggiungimento del ruolo di Professore ordinario.
Entrambi i campioni di docenti considerati sono caratterizzati da informazioni note solo parzialmente: nel primo le durate di carriera sono censurate per tutti i docenti che raggiungono il ruolo di Professore ordinario prima dell'anno 2000; nel secondo non tutti i Ricercatori in esame sperimentano l'evento entro la fine del periodo osservazionale, risultando quindi censurati. I metodi di analisi impiegati utilizzano come variabile risposta le durate di carriera, calcolate in anni a partire dalla prima pubblicazione di ogni docente.
I dati necessari per lo svolgimento di tale tesi sono stati reperiti in maniera autonoma e aggregati al fine di costruire i due dataset di interesse. Nel primo Capitolo si definiscono gli indici bibliometrici utilizzati e le relative informazioni necessarie alla loro costruzione. Nel secondo Capitolo viene descritta la metodologia utilizzata per l'estrazione dei dati dal *Web*, l'unione dei dati per la costruzione del dataset finale e il relativo calcolo degli indici nella maniera desiderata. Successivamente si mostrano una serie di analisi grafico-esplorative sugli indici scelti, motivando la decisione di considerarne solo due tra i quattro mostrati. Le analisi dei due campioni vengono separate nei due capitoli successivi. Nel quarto Capitolo vengono adattati diversi modelli per i dati di durata, focalizzando l'attenzione sul campione scelto a partire dai Professori ordinari al 31/12/2021. Viene utilizzato lo stimatore di Kaplan-Meier per valutare se gli indici, a

diversi stadi di carriera, abbiano un impatto diverso sul rischio di sperimentare l'evento. I metodi implementati attraverso il modello di Cox vengono tra loro confrontati utilizzando la log-verosimiglianza predittiva e i criteri di selezione automatica. Disponendo delle 'storie accademiche' e 'bibliografiche' di ciascun docente, di anno in anno, si sono inoltre utilizzati gli indici bibliometrici come covariate dipendenti dal tempo, mettendo a confronto il modello di Cox e il modello GAMM. Infine, attraverso le foreste casuali di sopravvivenza si sono confrontati i modelli stimati utilizzando l'errore di previsione. Nel quinto Capitolo si considera il campione di Ricercatori al 31/12/2005: dopo aver applicato il modello di Cox, viene effettuata una classificazione su diversi orizzonti predittivi considerando gli indici a uno, cinque, dieci anni da inizio carriera.

Capitolo 1

Analisi bibliometrica: impatto scientifico di un autore

Tra gli anni '60 e '80 dello scorso secolo è nata la scientometria, la scienza che si occupa della misurazione e dell'analisi delle pubblicazioni scientifiche. Valutare la ricerca significa stabilire norme e criteri per esprimere giudizi sulla qualità di una produzione e per misurarne la quantità. Come tecnica ha ampie applicazioni: comprendere la struttura di una disciplina, identificare i trend e le reti di ricerca, calcolare l'impatto scientifico di una pubblicazione, di singoli autori e gruppi di ricerca. Le istituzioni di ricerca utilizzano la scientometria come strumento decisionale per l'allocazione di fondi, decisioni politiche, per il confronto di autori di diversa provenienza.

1.1 Analisi bibliometrica

Gli indicatori di impatto scientifico possono essere definiti come 'strumenti statistici' atti a misurare gli aspetti quantificabili di creazione e diffusione del sapere scientifico. Nonostante l'impatto delle riviste sia tradizionalmente l'aspetto più valutato negli studi bibliometrici, anche l'interesse nei confronti della ricerca di singoli ricercatori, di recente, ha avuto un forte impatto.

A tale scopo sono stati sviluppati approcci per valutare la ricerca sia quantitativamente che qualitativamente. Tra questi, l'analisi bibliometrica nasce dal presupposto che la produzione di un ricercatore abbia valore solo quando è sottoposta al giudizio di un

comitato di individui aventi merito scientifico. La bibliometria è un metodo di analisi quantitativa che utilizza strumenti matematici e statistici al fine di misurare l'interrelazione e gli impatti delle pubblicazioni all'interno di un'area di ricerca. L'idea di base è quella di misurare il sapere scientifico attraverso l'analisi quantitativa di una qualsiasi unità di comunicazione (sia essa un articolo, un capitolo di un volume, un *paper*): una ricerca, dopo essere stata pubblicata su una rivista e sottoposta a 'referaggio' da parte di esperti, sarà citata da altri che la utilizzeranno come spunto per le loro successive produzioni. Diversi indicatori sono stati sviluppati negli anni con lo scopo di misurare l'impatto di singoli autori o di riviste. Vi sono metriche di carattere diverso che tengono conto di un numero variabile di fattori: il numero di citazioni totali, la distribuzione tra articoli o riviste, il numero medio di citazioni per pubblicazione, l'impatto di una rivista.

1.2 Indici bibliometrici a livello di autore

Nel presente lavoro risulta di interesse valutare metriche a livello di autore. Nel seguito verranno mostrate quelle utilizzate.

1.2.1 Indice di Hirsch

Uno degli indici più recenti e di successo è stato proposto dal fisico Hirsch (2005) per quantificare la prolificità di un autore e l'impatto delle sue pubblicazioni. Comunemente chiamato *h-index* dall'inglese è un indicatore della performance individuale dei singoli ricercatori, ma applicabile anche a gruppi di ricerca e istituzioni. L'originale definizione proposta dallo stesso Hirsch è stata:

> «*A scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq$ h citations each.*»

Pertanto un autore con *h-index* $= 8$, ad esempio, ha prodotto otto lavori ciascuno citato almeno otto volte.

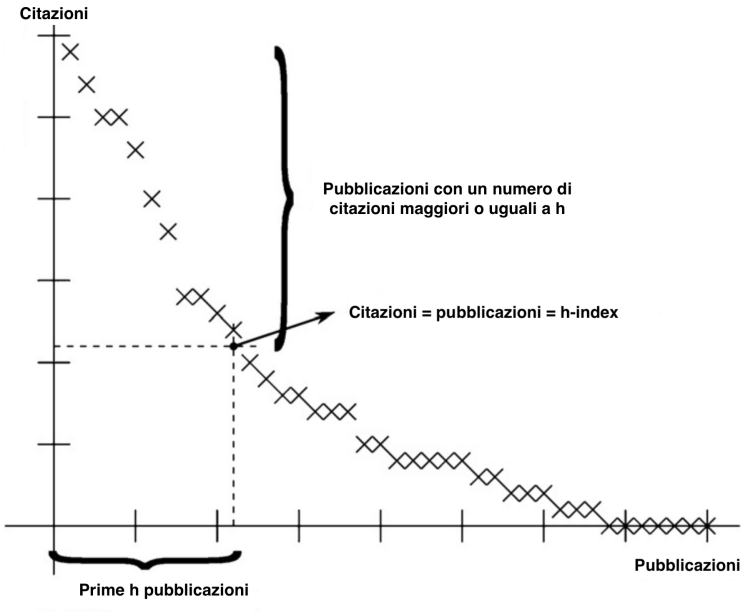

Figura 1.1: Interpretazione grafica dell'*h-index*: sull'asse delle ascisse sono riportate le pubblicazioni in maniera decrescente per numero di citazioni, dalla più alla meno citata. L'indice h corrisponde al valore della coordinata in cui il numero di citazioni (asse delle ordinate) equivale al numero di pubblicazioni.

Come mostrato in Figura 1.1, l'*h-index* è calcolato tenendo conto della quantità di pubblicazioni e del numero di citazioni ricevute per ognuna. Si tratta di una metrica che valuta la produttività scientifica di un autore in maniera cumulata. Esso fornisce una stima dell'impatto che un ricercatore ha, confrontandone la quantità con la qualità, in termini di pubblicazioni e citazioni, ordinando le pubblicazioni per il 'numero di volte citate'. L'indice è costruito per dare una valutazione completa sull'intera carriera di un autore e non per uno specifico intervallo di tempo. Per questo, ai fini del calcolo, Hirsch suggerisce di utilizzare le pubblicazioni tratte dalle intere carriere di un autore. L'*h-index* può variare a seconda della banca dati bibliografica o del motore di ricerca da cui viene estratto, a causa della diversa copertura bibliografica e temporale, nonché per i possibili errori legati a casi di omonimia e omografia.

Hirsch sostiene che l'*h-index* sia preferibile rispetto ad altri criteri basati su singoli numeri, come ad esempio il numero totale di pubblicazioni, il numero totale di citazioni o le citazioni per ciascuna pubblicazione. Tuttavia egli stesso ne sottolinea alcuni limiti. Un singolo numero non può dare che un'approssimazione grezza della produttività scien-

tifica di un autore, dunque sarebbe necessario considerare ulteriori fattori in maniera combinata. Oltre al problema dell'omonimia e dell'incompletezza nelle banche dati bibliografiche si riscontrano particolari differenze tra i diversi campi di ricerca. In particolare tra le diverse aree scientifiche l'indice potrebbe differire a causa del numero medio di autori partecipanti a una pubblicazione o al numero medio di paper pubblicati da ognuno. Si sottolinea però che tale limitazione è condivisa da qualsiasi indicatore basato sul conteggio delle citazioni, date le differenze tra discipline in termini di pratiche citazionali.

Vanclay (2007) ha posto l'attenzione su un importante beneficio dell'indice: la robustezza rispetto ad errori nella registrazione delle citazioni. L'*h-index* non è sensibile ai *paper* con poche citazioni e ciò che si è riscontrato negli anni è che la maggior parte degli errori nelle banche dati citazionali si ha nelle pubblicazioni con un basso numero di citazioni, le quali non influiscono nel calcolo di h. Tuttavia, allo stesso tempo, tale aspetto risulta essere negativo. Visto che solo le pubblicazioni più citate sono rilevanti nel calcolo, è irrilevante conoscerne il numero esatto. Tale caratteristica è opposta a quella usata dai criteri di selezione degli indicatori di impatto scientifico. I ricercatori con *paper* molto citati potrebbero avere un valore dell'indice simile o uguale a quelli con *paper* con un numero moderato di citazioni. Per un autore con un h estremamente piccolo e un numero di pubblicazioni limitato, ma un elevato numero di citazioni, l'indice h non riflette a pieno il valore. Inoltre la formula originale dell'*h-index* tiene in considerazione le auto-citazioni e fenomeni quali le citazioni 'gratuite' tra colleghi, le quali possono distorcere il valore dell'indice. Infine aspetti ignorati dall'indice sono il ruolo degli autori, il tipo di pubblicazione, le caratteristiche dei coautori presenti.
Come accade per altri indicatori, l'*h-index* dipende dalla durata di carriera di ciascun autore. Poiché il numero di pubblicazioni e citazioni aumenta nel tempo l'indice non può essere considerato una metrica universale in quanto non risulta adeguato per il confronto di autori a diversi stadi di carriera.
L'indice è stato utilizzato in letteratura per diversi scopi e in differenti contesti, tra i quali: il confronto diretto delle produzioni dei ricercatori, per misurare l'output complessivo di un gruppo completo di ricerca, istituzioni o gruppi di autori, per la valutazione dell'impatto scientifico di riviste, per confrontare le prestazioni di ricerca tra diversi

paesi.

1.2.2 Indice Universale

Un indice proposto di recente, nato dalla collaborazione dei membri del comitato dell'USERN (*Universal Scientific Education and Research Network*), sviluppato per porre rimedio ad alcune inefficienze dei preesistenti è l'*Universal Index (U-Index)* (Asaolu et al., 2022). Diversamente dall'*h-index* e da altri indici da esso derivati esistenti in letteratura, quest'ultimo tiene conto di informazioni quali la differenza tra l'autore principale e i coautori di una pubblicazione, la quale risulta essere un'importante caratteristica dei meriti di un autore nella maggior parte delle discipline. Inoltre tiene conto delle pubblicazioni totali di un autore e del tipo di pubblicazione, elementi tralasciati ad esempio nell'*h-index* e causa di sovrastima o sottostima della qualità di ricerca di un autore.

La scelta di includere tali criteri è avvenuta tramite le risposte raccolte da un questionario sottoposto ai membri del comitato dell'USERN, che include l'1% dei migliori scienziati provenienti da 21 campi scientifici differenti.
In Tabella 1.1 si riportano le risposte dei partecipanti (tasso di risposta del questionario 88%) riguardo la scelta delle variabili da includere nella costruzione dell'indice:

Parametri	**Frequenza (%)**
D'accordo con l'inclusione del tipo di pubblicazione	44 (100)
D'accordo con l'inclusione del numero di citazioni	43 (97.7)
D'accordo con l'inclusione della metrica relativa all'impatto della rivista	42 (95.5)
D'accordo con l'inclusione del numero di coautori	36 (81.8)
D'accordo con l'inclusione del ruolo dell'autore	34 (77.3)
Risposte totali	44

Tabella 1.1: Questionario per la scelta delle informazioni rilevanti da includere nella costruzione dell'*U-index*

Sono stati quindi assegnati specifici pesi a ciascuna variabile inclusa nella costruzione dell'indice u. Il contributo di ciascun parametro al punteggio totale è il seguente: 30% per il numero di citazioni, 20% per il tipo di pubblicazione, 20% per la misura di ciascuna rivista, 20% per il ruolo dell'autore, 10% per il conteggio dei coautori. Aggregando questi 5 parametri attraverso un'opportuna formulazione, ogni pubblicazione riceve un

punteggio definito *Universal Score for Publication* (USP), che varia da 0.1 a 1. Ciascun parametro, relativo ad ogni singola pubblicazione, riceve i seguenti sottopunteggi:

- numero di citazioni: [0, 0.3);
- tipo di pubblicazione: {0.1, 0.2};
- impatto della rivista: [0, 0.2);
- ruolo dell'autore: (0, 0.2];
- conteggio dei coautori: (0, 0.1].

L'indice proposto include parametri ignorati da altri indicatori a livello di autore. Tuttavia, come per l'*h-index*, esso non omette le auto citazioni e non considera le citazioni nascoste che includono citazioni nei documenti supplementari, i quali spesso mancano nelle banche dati citazionali. Inoltre potrebbe essere presente una distorsione nei parametri utilizzati per il calcolo, della quale non si tiene conto, data ad esempio dall'età, il genere, la classe sociale, la provenienza degli autori coinvolti.

1.2.3 *Citing author index* e *m-quotient*

Si menzionano inoltre in questo lavoro due ulteriori indici, l'*m-quotient* e il *citing authors index*, mostrati a solo titolo descrittivo.

Il primo nasce da alcune delle critiche mosse nei confronti dell'*h-index*, in particolare a riguardo del tempo e del numero di autori. In accordo con un modello stocastico per il processo di citazioni/produzione di un autore, Burrell (2007) sostiene che l'*h-index* sia approssimativamente proporzionale alla lunghezza di carriera. Un miglioramento dell'*h-index*, visto che non risulta adeguato per il confronto di ricercatori a diversi stadi di carriera, può essere ottenuto dividendo l'indice stesso per il numero di anni di attività di ricerca. Per questa ragione, Hirsch (2005), ha proposto di dividere l'indice h per il numero di anni di carriera dalla prima pubblicazione e così definire l'*m-index*.

Il secondo proposto, diversamente dai precedenti, focalizza l'attenzione sul numero di autori che hanno citato una pubblicazione, invece che sul numero di citazioni ricevute. L'idea di prendere in considerazione il numero di autori che citano ciascuna

pubblicazione è stata inizialmente introdotta da Dieks and Chang (1976), i quali hanno definito l'atto del citare come un processo probabilistico, e di recente Ajiferuke et al. (2010) hanno esteso l'idea alla valutazione dell'impatto della ricerca. È stato inoltre dimostrato che esiste una buona correlazione tra il *ca-* e il numero di citazioni totali, così come con l'*h-index* stesso (Cappelletti-Montano et al., 2021). Il *ca-index(r)* misura dunque l'impatto dell'autore r in termini di numero di distinti autori, esclusi i coautori e l'autore stesso, che hanno citato le pubblicazioni effettuate da r.

Capitolo 2

Indici bibliometrici nelle carriere accademiche

Le metriche descritte nel Capitolo 1 aiutano a tracciare l'impatto scientifico di un autore. Volendo mettere in relazione tali indici con le carriere accademiche dei docenti è sorta la necessità di costruire appositamente un dataset che aggregasse tali informazioni. Il processo di costruzione è avvenuto in fasi successive, avendo la necessità di estrarre informazioni da più siti e aggregarle. Nel seguito del capitolo verrà mostrato da dove si sono reperiti i dati e la relativa scelta del campione di interesse, il procedimento utilizzato per ottenerli e il calcolo degli stessi indici; infine verrà mostrato come è avvenuta l'aggregazione dei dati in un unico dataset. L'intero procedimento è avvenuto attraverso il supporto del software R.

2.1 Scelta del campione

Reperire informazioni sugli Accademici non risulta di così facile accesso in tutti i paesi: avere a disposizione le liste di docenti operanti nel sistema universitario non è sempre possibile. Per tale motivo si è infatti scelto di concentrarsi sull'Italia, in quanto grazie a *Cineca*, un consorzio interuniversitario senza scopo di lucro, cui aderiscono 69 università italiane, il Ministero dell'Istruzione e il Ministero dell'università e della ricerca oltre a 27 Istituzioni pubbliche Nazionali (Enti di ricerca, Aziende Ospedaliere Universitarie, ecc.) è possibile accedere liberamente alle liste di Ricercatori, Associati e Professori

Ordinari in servizio.
Il sito di *Cineca* mette a disposizione il servizio 'Cerca Università', all'interno del quale i docenti possono essere ricercati tramite 'Nome e Cognome'. È inoltre possibile specificare altri parametri di ricerca, tra i quali l'anno di operatività al quale ci si vuole riferire. Tuttavia il servizio offerto è limitato in termini temporali, in quanto la situazione accademica ricercabile è possibile all'indietro soltanto fino al 31/12/2000. Tale aspetto ha dunque vincolato la scelta del campione. Al fine di poter osservare l'andamento dei docenti nel tempo, in concomitanza alle loro carriere accademiche, si è quindi pensato di utilizzare come campione un gruppo di Professori Ordinari al 31/12/2021 e un gruppo di Ricercatori al 31/12/2005. In questo modo, la scelta di tali campioni, permette di osservare per uno specifico intervallo di tempo gli andamenti delle carriere di ciascuno, rispettivamente retrospettivamente e prospetticamente.

2.1.1 Cineca

L'attività principale di *Cineca* è il supporto alle attività di ricerca della comunità scientifica accademica; inoltre fornisce servizi di calcolo alle università in Italia. La collaborazione con il Ministero della pubblica istruzione ha inizio nel 1984, per la gestione dell'allora concorso nazionale a Professore ordinario. In seguito con il Ministero dell'istruzione, dell'università e della ricerca, il *Cineca* ha sviluppato diversi sistemi, diventandone il principale erogatore di servizi per quanto concerne le università. Esso aggrega diverse ricerche nell'ambito accademico, dai corsi di laurea ai finanziamenti, dai docenti agli studenti universitari, dai bandi alle statistiche. Alcune ricerche basilari sono possibili già dall'*homepage*, altre si possono effettuare dalle ricerche avanzate o da siti dedicati.
Nel presente lavoro il servizio 'Ricerca avanzata' è stato utilizzato per estrarre i ruoli accademici occupati negli anni dai docenti selezionati nei due campioni. Oltre alla possibilità di ricerca tramite 'Nome e Cognome', si può individuare un gruppo di docenti condizionando la scelta a una serie di fattori, tra i quali il 'Ruolo', il 'Settore', l''Area di ricerca' e il 'Macrosettore' di appartenenza. I campioni di interesse si sono ottenuti, selezionando il 'Ruolo' di Professore ordinario al 31/12/2021 e di Ricercatore al 31/12/2005, rispettivamente. Per entrambi si sono considerati i 'Settori' SECS-S da 01 a 05, MAT da 01 a 06 e 08, GEO da 01 a 09, per l''Area di ricerca' 13, 01 e 04 e

'Macrosettore' 13/D, 01/A e 04/A. Il primo campione individuato è caratterizzato da 964 docenti, il secondo da 1158 docenti.

I Professori ordinari al 31/12/2021, selezionati nei settori di interesse, sono così ripartiti nelle tre 'Aree di ricerca':

- 251 docenti dell'Area 13;
- 514 docenti dell'Area 01;
- 199 docenti dell'Area 04;

I Ricercatori al 31/12/2005, selezionati nei settori di interesse, sono così ripartiti nelle tre 'Aree di ricerca':

- 195 ricercatori dell'Area 13;
- 624 ricercatori dell'Area 01;
- 339 ricercatori dell'Area 04;

Selezionando ad esempio i Professori Ordinari, del 'Settore' SECS-S/01, per l''Area' 13 e 'Macrosettore' 13/D, specificando la situazione al 31/12/2021, in Tabella 2.1 si riportano le informazioni reperite da *Cineca* per alcuni docenti:

Fascia	**Cognome Nome**	**Genere**	**Ateneo**	**Facoltà**	**S.C.**	**Struttura di afferenza**
Ordinario	AGOSTINELLI Claudio	M	TRENTO	-	13/D1	Matematica
Ordinario	ALFÒ Marco	M	ROMA 'La Sapienza'	-	13/D1	Scienze statistiche
Ordinario	ALLEVA Giorgio	M	ROMA 'La Sapienza'	-	13/D1	Metodi e modelli per l'economia, il territorio e la finanza

Tabella 2.1: Professori ordinari al 31/12/2021, 'Settore' SECS-S/01, 'Area' 13, 'Macrosettore' 13/D

Le variabili relative all''Ateneo', alla 'Facoltà' e alla 'Struttura di afferenza' non si sono mantenute in quanto non risultano di interesse per le analisi e solo considerando i ruoli al 31/12/2021 il 20% dei docenti non dispone di tali informazioni.

Selezionando ad esempio i Ricercatori, del 'Settore' SECS-S/01, per l''Area' 13, specificando la situazione al 31/12/2005 si ottengono le informazioni contenute in Tabella 2.2:

Fascia	Cognome Nome	Genere	Ateneo	Facoltà	S.C.	Struttura di afferenza
Ricercatore	AMATO Ester	F	GENOVA	Economia	-	Non disponibile
Ricercatore	ANDREANO Maria Simona	F	CHIETI-PESCARA	Economia	-	Economia
Ricercatore	BALSAMO Giuseppa	F	PALERMO	Economia	-	Scienze statistiche e matematiche

Tabella 2.2: Ricercatori al 31/12/2005 del 'Settore' SECS-S/01, per l''Area' 13

Si precisa che i filtri di ricerca su 'Settore concorsuale'(S.C.) sono applicabili solo per gli anni 2011 e successivi, per questo la colonna relativa a S.C. in Tabella 2.2 risulta vuota. Anche in questo caso soltanto considerando i ruoli al 31/12/ 2005, per il 42% dei Ricercatori le informazioni relative all''Ateneo', alla 'Facoltà' e alla 'Struttura di afferenza' sono mancanti: non si sono perciò mantenute tali variabili.

A partire dai due campioni selezionati, si sono successivamente reperite le informazioni relative alle posizioni accademiche occupate per ogni anno di carriera. Per i Professori ordinari al 31/12/2021, procedendo in maniera retrospettiva nel tempo sino al 31/12/2000 si sono rilevati i ruoli accademici occupati per ogni anno: la data di rilevazione non è modificabile in quanto l'aggiornamento della situazione su *Cineca* è fornito soltanto al 31/12. Dal campione di 964 docenti ne sono stati immediatamente esclusi 148 in quanto, ricercando retrospettivamente le informazioni relative alle loro carriere accademiche nel corso degli anni, essi non erano presenti nel database di *Cineca* fino all'anno 2000. Il numero totale di Professori ordinari considerati al 31/12/2021 è dunque pari a 816. Il dataset ottenuto, contenente su ciascuna riga il 'Ruolo' occupato al 31/12 di ogni anno, è dunque caratterizzato da 17 952 osservazioni.

In maniera analoga per i Ricercatori al 31/12/2005, prospetticamente, si sono rilevate le posizioni accademiche occupate sino al 31/12/2021. Dal campione di 1158 docenti ne sono stati immediatamente esclusi 222 in quanto, ricercando le informazioni relative alle carriere accademiche nel corso degli anni, essi non erano presenti nel database di *Cineca* fino al termine del periodo osservazionale. Il numero totale di Ricercatori considerati al 31/12/2005 è dunque pari a 936. Il dataset contenente su ciascuna riga il 'Ruolo' occupato al 31/12 di ogni anno, è dunque caratterizzato da 15 912 osservazioni.

2.2 Database citazionali

Le informazioni sulle produzioni scientifiche, documenti e pubblicazioni con relative citazioni possono essere reperite all'interno dei database citazionali.
I database citazionali sono raccolte di libri, *paper*, articoli e altro materiale, il quale entra in un sistema *online* in un modo strutturato. Le informazioni relative a un singolo documento contenuto nel database (autore, titolo, dettagli della pubblicazione, *abstract*, l'intero testo) rappresentano il 'record' per quel documento. Ciascuna di queste caratteristiche diventa un 'campo' separato per quel record, permettendo di ricercare il documento stesso all'interno del database.
Quando un documento entra originariamente nel database è analizzato attraverso le parole chiavi che lo caratterizzano e viene di conseguenza assegnato a categorie precise che permettono successivamente di individuarlo. Tuttavia l'accesso a tali fonti di dati non è sempre liberamente possibile.
Le due banche dati citazionali a livello internazionale sono *Clarivate Analytics' Web of Science* (WoS) e *Scopus* di Elsevier, alle quali si aggiunge una molto diffusa e liberamente accessibile, *Google Scholar*. Ciascuna di esse contiene informazioni di carattere diverso e in particolare la copertura in termini di dati contenuti risulta differente. Secondo numerosi studi (Yang and Meho, 2007) i primi due citati assicurano un'elevata qualità e affidabilità dei dati, mentre l'ultimo assicura un grado di copertura maggiore, ma una bassa qualità e affidabilità a livello di contenuti.
Con lo sviluppo e continuo aggiornamento di tali banche dati citazionali, è stato possibile mettere in relazione tra loro pubblicazioni, conoscerne il numero di citazioni, valutare la produttività e l'impatto degli autori.

2.2.1 *Scopus: web scraping*

Tra i database citati, nel presente lavoro si è scelto di utilizzare *Scopus* come fonte di dati. *Scopus* è un database creato nel 2004 dalla casa editrice Elsevier: periodicamente aggiornato, esso contiene informazioni su più di 37 000 articoli e su più di 12 000 autori. La copertura temporale dei dati in esso contenuti, seppure parziale, si estende fino all'anno 1956. Il portale permette la visualizzazione di una serie di dati, tra i quali *abstract*, articoli, pubblicazioni suddivise per autori, istituzioni, riviste, ecc. Tuttavia senza una sottoscrizione o d'istituto o personale le informazioni che si possono reperire

sono soltanto quelle di base, relative agli autori. Dopo aver effettuato la sottoscrizione al sito, l'accesso permette di consultare e scaricare dati di diverso tipo, manualmente o in maniera automatica, effettuando *web scraping*.
La motivazione per la quale si è preferito *Scopus* rispetto ad altre fonti è legata principalmente alle informazioni riguardanti ogni singola pubblicazione in esso contenuta. Poiché l'obiettivo nel presente contesto è quello di voler costruire gli indici bibliometrici per ciascun autore, in modo tale da avere una misura di produttività aggiornata di anno in anno, le informazioni reperite su *Scopus* sono esaustive a tale fine. Effettuando la ricerca attraverso 'Nome e Cognome', il database fornisce per ciascun autore la lista di pubblicazioni e per ognuna di queste le seguenti informazioni:

- Ordine dell'autore nella pubblicazione
- Numero di citazioni della pubblicazione
- Numero di coautori della pubblicazione
- Titolo della rivista su cui è stata pubblicata
- Tipologia di pubblicazione: articolo, *paper*, ecc.
- Data della pubblicazione

Visto il presente contesto nel quale si vogliono ottenere le pubblicazioni relative a un numero elevato di docenti, farlo manualmente avrebbe richiesto uno sforzo oneroso. Tuttavia indipendentemente da come viene effettuata l'operazione di estrazione, per ogni ricerca attraverso un 'campo' (nome dell'autore, pubblicazioni su una rivista, ecc.) i dati che si possono scaricare non sono infiniti. Poter ottenere una grande quantità di informazioni (più di 20 000 'record') è comunque possibile soltanto attraverso il possesso di un *API-key*, che permette di scaricare i dati in maniera automatica effettuando *web scraping*. La chiave di accesso personale dev'essere richiesta ai realizzatori del sito. Tuttavia quest'ultima è concessa sotto opportune condizioni e motivando gli scopi che si intendono perseguire con i dati in questione.

Nell'operazione di estrazione automatica, è possibile ricercare i documenti di un singolo autore mediante 'Cognome Nome' o attraverso lo 'Scopus ID' (identificativo numerico di un autore). Per ovviare a problemi di omonimia, frequenti in questi casi, ci

si è avvalsi dell'identificativo di ciascun docente, ricercandolo nel sito stesso. L'operazione di estrazione automatica consente di ottenere tutte le informazioni di interesse in maniera già strutturata all'interno di un dataset. La Tabella 2.3 riporta alcune delle informazioni estratte per un autore:

	auth-order	n-auth	citations	journal	description	cover-date	au-id
1	3	3	4	Journal of Computational and Graphical Statistics	Article	2022-01-01	23391969800
2	2	2	0	Test	Article	2021-12-01	23391969800
3	2	3	3	Metron	Article	2021-06-01	23391969800
4	3	3	6	Statistical Methods and Applications	Article	2021-03-01	23391969800

Tabella 2.3: Struttura dataset estratto da *Scopus*

Ogni riga in Tabella 2.3 riporta le informazioni relative a ogni singola pubblicazione effettuata dall'autore in questione. Le variabili riportate, rispettivamente da sinistra a destra, corrispondono a: ordine dell'autore tra gli autori di una pubblicazione, numero di autori partecipanti ad una pubblicazione, numero di citazioni di una pubblicazione, giornale sulla quale è stata pubblicata, data relativa alla pubblicazione e identificativo dell'autore. Si sono tralasciate le informazioni relative all'affiliazione (Università, Ente, ecc.), al titolo della pubblicazione, all'*abstract* e alle parole chiave in quanto non rilevanti ai fini delle analisi.

Per alcune pubblicazioni le variabili estratte sono risultate incomplete: l'ordine dell'autore ('auth-order') e la tipologia di giornale ('description'). I dati mancanti sono stati imputati nel seguente modo:

- all'ordine mancante dell'autore si è assegnato il valore 'uno', ossia primo autore. La frequenza relativa di valori mancanti per il campione selezionato a partire dai Professori ordinari al 31/12/2021 è pari a 0.0005; per i Ricercatori al 31/12/2005 è pari a 0.0021.

- al tipo di giornale si è assegnato il valore *Article*, ossia quello più frequente. La frequenza relativa di valori mancanti per il campione selezionato a partire dai Professori ordinari al 31/12/2021 è pari a 0.0003; per i Ricercatori al 31/12/2005 è pari a 0.0001.

2.3 Unione delle informazioni: costruzione del dataset

Disponendo a questo punto delle due informazioni di interesse, i ruoli nelle carriere accademiche nel periodo considerato e le pubblicazioni nel corso delle carriere, il database reperito da *Cineca* e quello reperito da *Scopus* sono stati uniti attraverso la variabile comune 'Cognome Nome'. L'operazione è stata effettuata congiungendo i due database in maniera tale da mantenere lo stesso numero di righe del dataset ottenuto da *Scopus*. In questo modo, per ciascuna riga, si dispone delle informazioni relative ad ogni singola pubblicazione e al ruolo in carriera occupato dal docente nell'anno della pubblicazione in esame.

Sono state effettuate alcune operazioni sulle variabili originali estratte:

- la variabile 'Area di ricerca' è stata creata assegnando i docenti dell''Area 13, 01 e 04' rispettivamente a 'Statistica, Matematica e Geoscienze';
- la variabile 'Fascia' originariamente estratta da *Cineca* è stata cambiata in 'Ruolo accademico'. Essendo l'informazione relativa al ruolo nota a partire dall'anno 2000, il 'Ruolo accademico' caratterizzato da zeri indica che non si conosce quale posizione il docente ricoprisse in quel momento;
- la variabile 'Cover-Date' presente nel dataframe estratto da *Scopus* è stata modificata mantenendo solo l'anno di riferimento e cambiata in 'Anno'.

Poiché per entrambi i campioni selezionati, l'interesse è quello di valutare gli indici bibliometrici nelle carriere fino al momento in cui essi diventano Professori ordinari, si sono mantenute le pubblicazioni soltanto fino all'anno corrispondente a tale evento.

Per entrambi i dataset ottenuti la variabile risposta di interesse è rappresentata dalla durata di carriera impiegata da ciascuno per divenire Professore ordinario. Poiché né *Scopus*, né *Cineca* forniscono informazioni relative all'età o all'anno di inizio carriera di un docente, si è utilizzata come data di inizio carriera l'anno della prima pubblicazione reperita sul database citazionale. La durata di carriera è stata dunque calcolata come differenza tra l'anno in cui un docente diventa Professore ordinario e l'anno della prima pubblicazione. Si indica con Y_{i_k} per ciascuna riga dei due dataset, gli anni di

carriera trascorsi a partire dalla prima pubblicazione, per il docente i, nell'anno della pubblicazione k.

Nel campione di docenti scelto a partire dai Professori ordinari al 31/12/2021, si è osservato che per 15 di questi la prima pubblicazione disponibile è successiva all'anno in cui raggiungono il ruolo di Professori ordinari. Le osservazioni corrispondenti a tali soggetti sono state perciò eliminate.

I due dataset sono quindi così caratterizzati:

- $i = 1, ..., 801$ Professori ordinari al 31/12/2021; $k = 1, ..., 25\,634$ pubblicazioni;
- $i = 1, ..., 936$ Ricercatori al 31/12/2005; $k = 1, ..., 33\,729$ pubblicazioni;

2.4 Operazionalità degli indici bibliometrici

Volendo valutare l'andamento degli indici nel tempo, in concomitanza all'attività accademica di ciascun docente, è stato necessario calcolare i quattro indici definiti nel Capitolo 1, utilizzando le informazioni relative alle singole pubblicazioni.

2.4.1 *H-index*

L'*h-index* richiede le seguenti variabili per il calcolo:

- pubblicazioni effettuate da ogni docente;
- numero di citazioni ricevute per ogni pubblicazione.

Si consideri il docente i, il quale nel corso della carriera ha effettuato $k = 1, ..., l, ..., K$ pubblicazioni e ciascuna pubblicazione riceve un numero di citazioni $n = 0, ..., N$. L'indice h è stato calcolato nel seguente modo:

1. le K pubblicazioni vengono ordinate dalla meno alla più recente;
2. la prima pubblicazione $k = 1$ ha indice $h = k$, se ha ricevuto almeno $n >= k$ citazioni;
3. al considerare ogni pubblicazione successiva, $k = 2, 3, ..., l$, si riordinano le $k = 1, ..., l$ pubblicazioni in senso decrescente per il numero di citazioni n ricevute da ciascuna;

4. l'indice h viene ricalcolato in ciascuna riga, all'arrivo di ogni nuova pubblicazione l. L'ordine delle pubblicazioni $k = 1, ..., l$ varia se l'*l-esima* pubblicazione ha ricevuto un numero di citazioni più elevato rispetto alla pubblicazione $l - 1$;

5. contemporaneamente si tiene conto dell'informazione relativa all'indice h raggiunto ad ogni anno di carriera;

6. il dataset viene infine riordinato dalla pubblicazione meno recente alla più recente, disponendo così dell'indice aggiornato ad ogni anno di carriera del docente in esame.

2.4.2 *U-index*

I parametri utilizzati per il calcolo dell'*u-index* sono stati ottenuti a partire dalle informazioni reperite da *Scopus*. Per ciascuna pubblicazione si è tenuto conto di:

- C: numero di citazioni;

- T: tipo di pubblicazione. Viene assegnato il valore $T = 0.1$ per lettere, note, report, editoriali e $T = 0.2$ per articoli, revisioni, brevi sondaggi, riviste di dati, capitoli di libri e libri. Tali categorie sono state adottate dalla lista di tipologia di pubblicazioni presenti su *Scopus*;

- S: misura dell'impatto della rivista sulla quale viene pubblicata (qui il parametro è stato calcolato utilizzando la mediana annuale del *CiteScore*(CS)[1] di tutti i giornali presenti su *Scopus*);

- R: ruolo dell'autore. Nelle pubblicazioni con un numero di autori inferiore a 100 si considera $R = 1$ per il primo, ultimo e gli autori corrispondenti, e $R =$ 'posizione' (tra tutti gli autori nominati) per gli altri coautori. Nelle pubblicazioni con più di 100 autori, si è considerato analogamente $R = 1$ per il primo, l'ultimo e gli autori corrispondenti, mentre $R = 99$ per tutti gli altri coautori. Tale decisione si è basata sul fatto che un numero elevato di autori (>100) ha un effetto diretto sulla frequenza con cui le pubblicazioni vengono citate;

[1] Misura dell'impatto di una rivista fornita da Scopus, calcolato considerando gli ultimi quattro anni precedenti quello corrente.
Ad esempio per l'anno 2021, $CS_{2021} = \frac{Citazioni\ ricevute\ dalla\ rivista\ tra\ il\ 2018\ e\ il\ 2021}{Documenti\ pubblicati\ nella\ rivista\ tra\ il\ 2018\ e\ il\ 2021}$

- N: numero di coautori. Si è considerato $N = 1$ per il primo, l'ultimo e gli autori corrispondenti.

Il calcolo finale dell'indice, considerando tutte le $k = 1, ..., K$ pubblicazioni per il docente i dalla meno alla più recente, avviene attraverso la seguente formula:

$$U\text{-}index = \sum_{k=1}^{K} USP_k \times E_i(USP) \tag{2.1}$$

dove

$$USP_k = T_k + \frac{0.2}{1 + \log(R_k)} + \frac{0.1}{1 + \log(N_k)} + \frac{(0.2 \times S_k)}{1.6 + S_k} + \frac{(0.2 \times C_k)}{50 + C_k} \tag{2.2}$$

sono i sottopunteggi calcolati per ogni pubblicazione k.

Si sottolinea che le pubblicazioni della tipologia 'Correzioni' o 'Errori di stampa' sono state escluse dal calcolo, in quanto ripetitive, così come non sono state considerate le pubblicazioni ritirate.

2.4.3 *Ca-index*

Il *ca-index* è definito come:

$$ca\text{-}index(i)_k = \max\{0, N_{a_k}(i) - N_{co_k}(i)\} \tag{2.3}$$

dove $N_{a_k}(i)$ rappresenta il numero di autori (inclusi eventuali possibili coautori) che hanno citato il *paper* k pubblicato da i e $N_{co_k}(i)$ è il numero di distinti coautori di i (incluso i) che hanno partecipato alla k-*esima* pubblicazione. La scelta di evitare eventuali valori negativi non muta il significato dell'indice, in quanto valori pari a zero possono identificare basse prestazioni dell'autore in maniera adeguata.

Le informazioni per la costruzione di tale indice si sono ottenute dalle variabili estratte da *Scopus*, relative al numero di autori che citano ciascuna pubblicazione, e di co-autori partecipanti ad ogni pubblicazione. Al fine del calcolo, considerando il docente i e le relative $k = 1, ..., K$ pubblicazioni ordinate dalla meno alla più recente, il

procedimento utilizzato per tener conto della crescita dell'indice nel corso della carriera è il seguente:

1. per la pubblicazione $k = 1$ l'indice *ca* è calcolato come differenza tra il numero di autori che hanno citato la pubblicazione k e il numero di autori che hanno partecipato alla pubblicazione k;

2. per le successive pubblicazioni, viene cumulato per $k = 1, ...l$ il numero di autori citanti N_{a_k} e il numero di autori N_{co_k} che hanno partecipato alle k pubblicazioni del docente in questione;

3. il valore dell'indice in corrispondenza della pubblicazione k, aumenta o rimane costante a seconda dell'incremento subìto da N_{a_k} e da N_{co_k}.

2.4.4 *M-index*

L'*m-index*, essendo dipendente dalla struttura dell'indice h non necessita di essere calcolato in maniera cumulata, diversamente dagli altri tre. Dopo aver ordinato le $k = 1, ..., K$ pubblicazioni per ogni docente i dalla meno alla più recente, l'indice per ogni pubblicazione è stato calcolato dividendo il valore dell'*h-index* per il numero di anni di carriera trascorsi dalla prima pubblicazione. Per ciascuna pubblicazione k il valore ottenuto è il seguente:

$$m\text{-}index(k) = \frac{h\text{-}index(k)}{Y_{i_k}} \tag{2.4}$$

dove $h\text{-}index(k)$ è il valore dell'indice h sulla riga corrispondente alla *k-esima* pubblicazione e Y_{i_k} il numero di anni di carriera del docente i trascorsi dalla prima alla *k-esima* pubblicazione.

2.5 Dataset finale

Dopo aver effettuato il calcolo degli indici considerando tutte le pubblicazioni estratte da *Scopus*, per ciascun docente, sono state eliminate le variabili necessarie alla costruzione degli stessi. Poiché la variabile di interesse nel presente lavoro è caratterizzata dal numero di anni di carriera impiegati per raggiungere il ruolo di Professore ordinario, si

è deciso di mantenere per entrambi i dataset una sola osservazione per anno. In particolare, avendo estratto l'aggiornamento relativo alla posizione in carriera al 31/12 di ogni anno e avendo calcolato gli indici riordinando le pubblicazioni di ciascun docente dalla meno alla più recente, l'osservazione mantenuta è quella corrispondente alla situazione registrata al termine dell'anno corrente. In questo modo il valore dell'indice calcolato per ogni anno considera l'intera produttività scientifica del docente in questione. I due dataset sono così costituiti:

- $k = 1, ..., 10\,732$ pubblicazioni per $i = 1, ..., 801$ Professori ordinari al 31/12/2021;
- $k = 1, ..., 14\,224$ pubblicazioni per $i = 1, ..., 936$ Ricercatori al 31/12/2005;

Le variabili presenti nei due dataset sono riportate in Tabella 2.4:

Variabile	**Descrizione**
Cognome Nome	Cognome, Nome dell'autore
Scopus ID	Identificativo *Scopus* del docente
Sesso	Maschio (M), Femmina (F)
Ca-index, U-index, H-index, M-index	Indici bibliometrici
Area di ricerca	Matematica, Geoscienze, Statistica
Ruolo	Ricercatore, Associato, Ordinario
Anno della pubblicazione	Anno a cui risale la pubblicazione
Anno di inizio carriera	Anno della prima pubblicazione
Anni di carriera	Differenza tra anno della pubblicazione in esame e anno della prima pubblicazione

Tabella 2.4: Descrizione delle variabili presenti nel dataset

In Tabella 2.5 sono riportate alcune righe e alcune delle variabili del dataset ottenuto a partire dal campione selezionato di Professori ordinari al 31/12/2021:

	Scopus ID	**Ca-index**	**U-index**	**H-index**	**M-index**	**Area di ricerca**	**Ruolo**	**Anno della pubblicazione**	**Anni di carriera**
1	10045163400	43	0.4	1	0	Matematica	0	1986	1
2	10045163400	43	0.74	2	2	Matematica	0	1987	2
...	...	...	...	...	...	...	...	...	...
8	10045163400	50	3.14	12	0.86	Matematica	Ordinario	2000	15
...	...	...	...	...	...	...	...	...	...
10 728	9338338400	0	0.33	1	0	Statistica	0	1988	1
...	...	...	...	...	...	...	...	...	...
10 732	9338338400	35	2.02	4	0.24	Statistica	Ordinario	2005	18

Tabella 2.5: Alcune righe del dataset selezionato a partire dal campione di Professori ordinari al 31/12/2021

La struttura del secondo dataset costruito, a partire dal campione di Ricercatori al 31/12/2005, è la medesima.

Capitolo 3

Analisi esplorative

In questo capitolo vengono mostrate alcune analisi grafico-esplorative su entrambi i dataset costruiti. Si considereranno separatamente le singole storie bibliografiche di alcuni docenti e gli andamenti collettivi degli indici per tutti i soggetti in esame.

3.1 Professori ordinari al 2021

3.1.1 Analisi preliminari

Si considerino le durate di carriera, dalla prima pubblicazione alla massima carica accademica, per ciascun docente presente nel campione di Professori ordinari al 31/12/2021. In Tabella 3.2 le distribuzioni delle durate di carriera suddivise per le tre 'Aree di ricerca':

Area disciplinare	**Minimo**	**1° quartile**	**Mediana**	**Media**	**3° quartile**	**Massimo**
Matematica	5.00	16.00	19.00	19.7	23.00	59.00
Geoscienze	3.00	20.00	24.00	23.42	26.50	43.00
Statistica	2.00	12.00	16.00	16.32	20.00	31.00

Tabella 3.1: Distribuzione del numero di anni di carriera, distinto per 'Area di ricerca'

Si osserva che i valori massimi delle durate di carriera sono estremamente elevati: impiegare 59 anni per raggiungere il ruolo di Professore ordinario è sostanzialmente impossibile. Allo stesso tempo un numero di anni di carriera inferiore a tre per divenire Professore ordinario risulta estremamente improbabile. Ricordando che gli anni di carriera sono stati calcolati a partire dalla prima pubblicazione reperita su *Scopus*,

valori molto piccoli o molto elevati di durate potrebbero essere dovuti a errori nella ricostruzione delle 'storie bibliografiche' sullo stesso sito. Le durate di carriera molto brevi potrebbero essere causate dal fatto che il database citazionale non contiene tutte le pubblicazioni di un autore, ma solo le più recenti. Le durate estremamente lunghe possono essere dovute a errate attribuzioni di pubblicazioni ad autori i quali non ne sono in realtà gli artefici. Ricercando su *Scopus* i docenti corrispondenti a tali durate anomale si è osservato infatti che le durate molto lunghe sono associate a persone con nomi molto frequenti nel database. Ai fini di ottenere risultati sensati nelle successive analisi si è perciò deciso di eliminare i docenti le cui caratteristiche non rispecchiano delle durate di carriera ritenute 'accettabili'. Il campione di 801 Professori ordinari al 31/12/2021 è stato perciò ridotto, mantenendo solo chi ha raggiunto la cattedra dopo i 5 ed entro i 30 anni dalla prima pubblicazione. Il dataset sul quale le analisi successive saranno effettuate è così caratterizzato:

- $i = 1, ..., 759$ docenti;
- $k = 1, ..., 10\,054$ pubblicazioni corrispondenti ai docenti in esame.

Poiché le informazioni relative ai ruoli accademici occupati sono note a partire dal 31/12/2000 in poi, le durate di carriera impiegate per raggiungere la posizione di Professore ordinario sono note con esattezza solo per chi raggiunge tale posizione a partire dal 31/12/2001. Tra i 759 docenti considerati, 122 risultano già Professori ordinari al 31/12/2000. I rimanenti 637 raggiungono il ruolo negli anni successivi considerati.

3.1.2 Analisi di singole bibliografie

Dal campione in esame si considerino due diversi docenti:

- il primo raggiunge il ruolo di Professore ordinario dopo il 31/12/2000, dunque la durata di carriera è nota con esattezza;
- il secondo al 31/12/2000 risulta già Professore ordinario. Per quest'ultimo la durata di carriera non è quindi nota con esattezza.

In Figura 3.1, 3.2, 3.3, 3.4 si riportano l'indice h, u, ca e m rispettivamente, per il primo docente considerato:

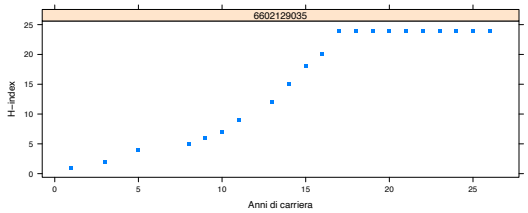

Figura 3.1: H-index

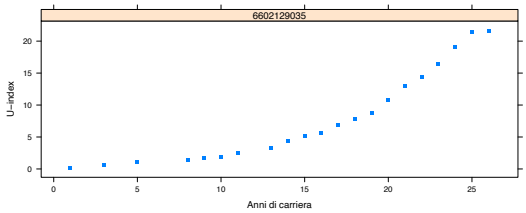

Figura 3.2: U-index

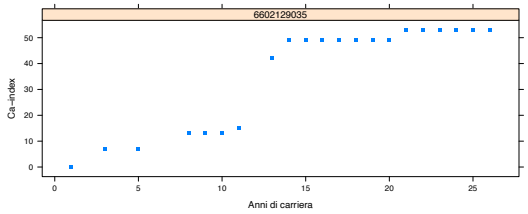

Figura 3.3: Ca-index

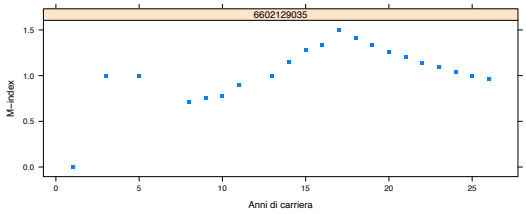

Figura 3.4: M-index

Il docente in esame raggiunge il ruolo di Professore ordinario al 26° anno di carriera. Dai grafici riportati si può osservare che l'indice h presenta un andamento crescente nel tempo fino al 17° anno di carriera, successivamente esso rimane costante. L'*m-index*, essendo dipendente da h, varia anch'esso fino al 17° anno di carriera: successivamente, all'aumentare degli anni di carriera e al valore costante di h, l'indice decresce. L'indice *ca* mostra un andamento inizialmente crescente nei primi anni di carriera e circa costante tra i 3 e i 12 anni. Un notevole aumento si nota in corrispondenza del 13° anno di carriera: questo potrebbe essere dovuto a un numero elevato di autori che hanno citato le pubblicazioni del docente in esame e alle quali hanno preso parte pochi autori. Diversamente dagli altri tre, l'indice u mostra un andamento sempre crescente nel tempo, dall'inizio di carriera al raggiungimento della cattedra.

Si consideri ora il secondo docente, il quale raggiunge il ruolo di Professore ordinario in un ignoto istante di tempo precedente il 31/12/2000. In Figura 3.5, 3.6, 3.7, 3.8 si riportano l'indice h, u, ca e m rispettivamente:

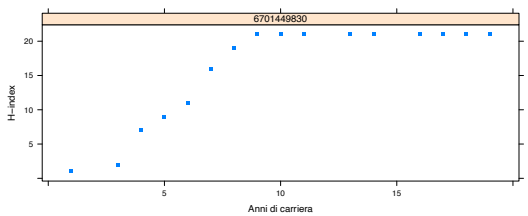

Figura 3.5: H-index

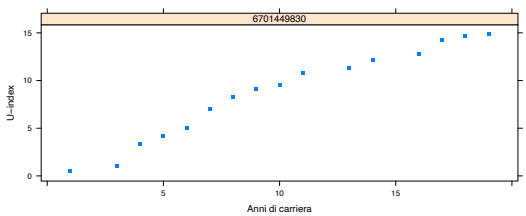

Figura 3.6: U-index

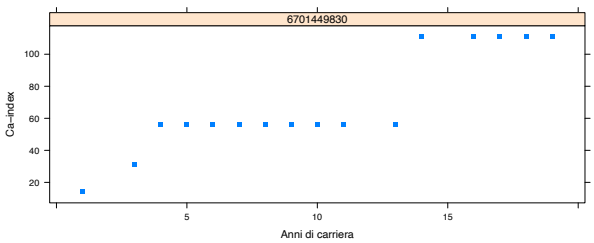

Figura 3.7: Ca-index

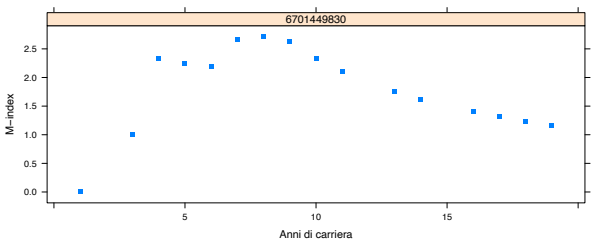

Figura 3.8: M-index

Il docente considerato diventa ordinario al 19° anno di carriera. Si consideri che tale durata non è nota con certezza, in quanto l'evento potrebbe essersi verificato con esattezza in corrispondenza dell'anno 2000 o in un qualche anno precedente. Si osserva innanzitutto che gli indici nel tempo assumono un comportamento simile al docente considerato in precedenza, nonostante i valori differiscano. Il docente qui considerato presenta valori degli indici h e u più bassi, rispetto al precedente e più alti per il *ca-* e *m-index*. In particolare si osserva che l'arresto della crescita dell'indice h avviene ancor prima rispetto al caso precedente, intorno all'8° anno di carriera. U e ca continuano invece a crescere notevolmente fino al 14° anno di carriera: negli anni successivi il primo si arresta, il secondo aumenta ancora lievemente. L'*m-index*, ancora una volta, mostra un effetto decrescente dal momento in cui l'*h-index* diventa costante.

Confrontando i due docenti, l'andamento crescente di h è più lento per il primo soggetto, il quale impiega un numero maggiore di anni per diventare Professore ordinario rispetto al secondo in questione. Anche il comportamento dell'indice u fornisce utili informazioni: il docente che impiega un numero di anni inferiore per raggiungere la massima carica mostra una crescita dell'indice, nei primi quindici anni di carriera, più rapida rispetto al primo considerato.

3.1.3 Analisi collettive

Utilizzando l'intero campione di docenti a disposizione si vuole ora osservare il comportamento degli indici a livello collettivo.

3.1.4 Indici bibliometrici in corrispondenza dell'evento di interesse

Si è osservato dai precedenti grafici individuali che ciascun indice utilizzato risente di un'unità di misura differente. La distribuzione dei valori, nell'anno in cui i docenti raggiungono il ruolo di Professore ordinario, è riportata in Tabella 3.2:

Indice	**Minimo**	**1° quartile**	**Mediana**	**Media**	**3° quartile**	**Massimo**
Ca-index	0.00	41.00	85.00	150.2	170.00	5247.00
U-index	0.46	4.59	8.21	10.38	13.84	83.85
H-index	1.00	9.00	14.00	16.34	21.00	67.00
M-index	0.05	0.57	0.83	0.98	1.21	4.43

Tabella 3.2: Distribuzione del numero di anni di carriera, distinto per Area di ricerca

Per i primi tre indici considerati le distribuzioni dei valori sono molto estese. Tuttavia si osserva che la mediana per l'indice *ca*, *u* e *h* si ha in corrispondenza dei valori 85, 8.2 e 14, rispettivamente. I valori più frequenti con i quali i docenti in questione raggiungono il ruolo di Professore ordinario non risultano quindi molto elevati.

In Figura 3.9, 3.10, 3.11, 3.12 si riportano le distribuzioni dei valori degli indici, nell'anno in cui i docenti raggiungono il ruolo di Professore ordinario, distinguendo le tre 'Aree di ricerca':

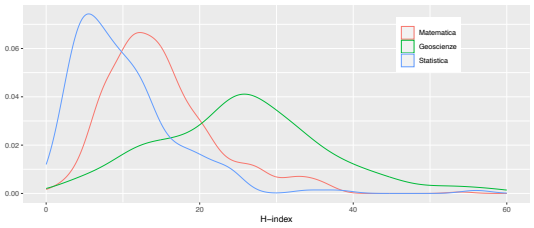

Figura 3.9: H-index

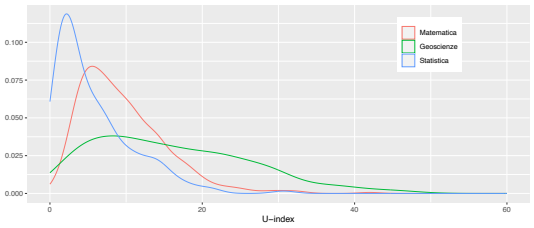

Figura 3.10: U-index

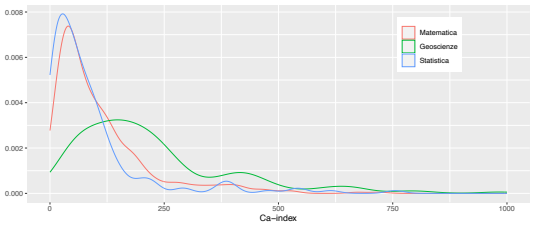

Figura 3.11: Ca-index

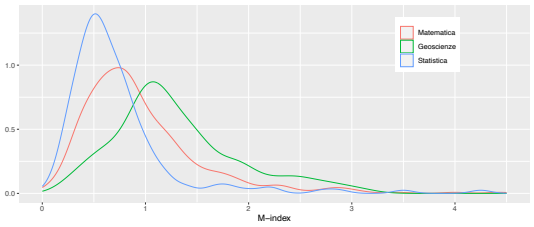

Figura 3.12: M-index

La categoria dei docenti di Geoscienze risulta avere, in generale per tutti gli indici, valori differenti rispetto alle altre due. In particolare per il *ca-* e l'*u-index* sono più frequenti valori bassi degli indici, al contrario per l'*h-* e l'*m-index*, i quali tendono ad

avere più frequentemente valori vicini alla media, compresi rispettivamente tra 20 e 30 per il primo e tra 1 e 1.5 per il secondo. Matematici e Statistici sembrano dimostrare un andamento delle frequenze dei valori più simile tra loro: in generale per tutti e quattro gli indici si nota come la distribuzione delle frequenze sia particolarmente asimmetrica a sinistra, concentrata sui valori più bassi.

3.1.5 Indici bibliometrici negli anni di carriera

Finora si è posta l'attenzione sui valori degli indici raggiunti in corrispondenza dell'anno in cui essi raggiungono il ruolo di Professore ordinario. Si vuole ora valutare più nel dettaglio come gli indici, negli anni di carriera, siano predittivi nel determinare le durate impiegate per raggiungere il massimo ruolo accademico.
Si osservano nelle Figure 3.13, 3.14, 3.15, 3.16 i boxplot contenenti le distribuzioni degli indici, all'aumentare degli anni di carriera, per le tre aree accademiche scelte:

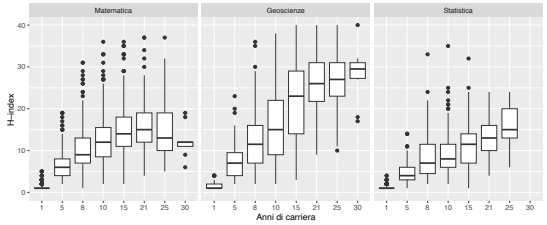

Figura 3.13: H-index

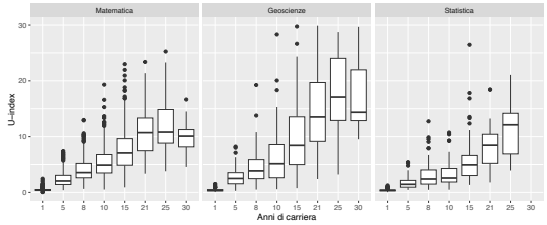

Figura 3.14: U-index

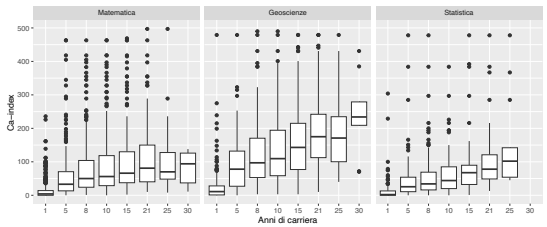

Figura 3.15: Ca-index

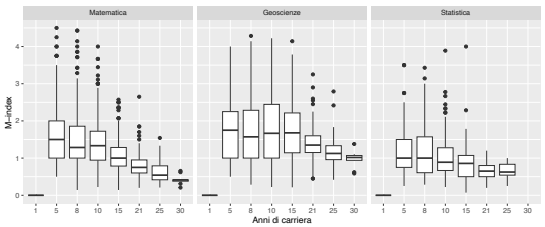

Figura 3.16: M-index

Come osservato nelle precedenti analisi individuali, l'andamento degli indici *ca*, *u* e *h* è crescente nel tempo, diversamente dall'*m-index* che risulta avere un andamento decrescente all'aumentare degli anni di carriera. Ciò si riscontra in quanto quest'ultimo è calcolato in termini di allontanamento dall'inizio di carriera: se l'*h-index* rimane costante al passare degli anni, l'*m-index* decresce velocemente. Tale effetto è notevole in particolare dopo i 10/15 anni dalla prima pubblicazione.
Si può inoltre osservare che gli indici per la categoria di Geoscienze assumono valori più

elevati, in media, rispetto alle altre due. I primi tre indici citati tendono a crescere più velocemente nel tempo e l'm a decrescere più lentamente. Poiché tale effetto è condiviso da tutti gli indici, la causa potrebbe essere dovuta a un fattore che li accomuna: il numero di citazioni più elevato rispetto ai docenti di Matematica e Statistica.

3.1.6 Indici bibliometrici tra Ricercatori e Associati

Si prendono in considerazione i valori degli indici u e h nelle carriere dei docenti. Per coloro che raggiungono la posizione di Professore ordinario in un tempo noto con esattezza, si riporta l'andamento degli indici in concomitanza con i ruoli accademici occupati in Figura 3.17, 3.18; per i docenti già ordinari al 31/12/2000 si mostra l'andamento degli indici nel tempo in 3.19, 3.20:

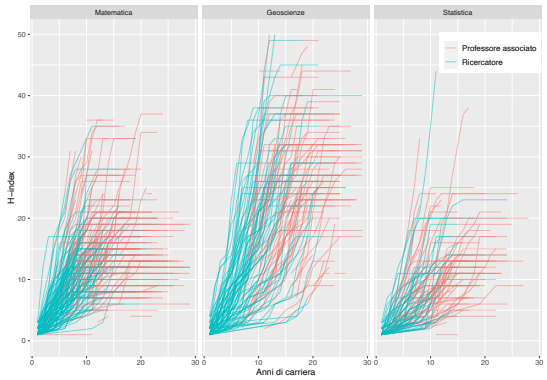

Figura 3.17: H-index

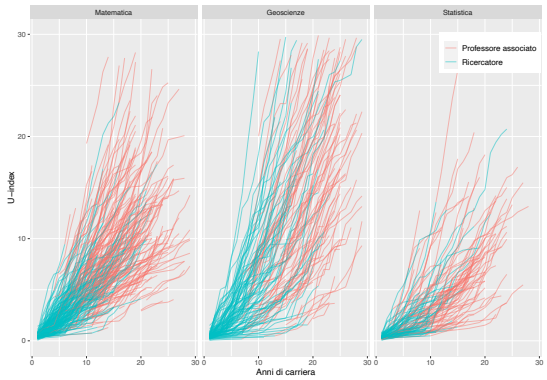

Figura 3.18: U-index

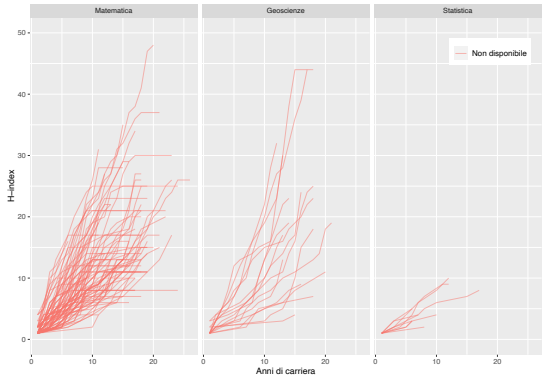

Figura 3.19: H-index

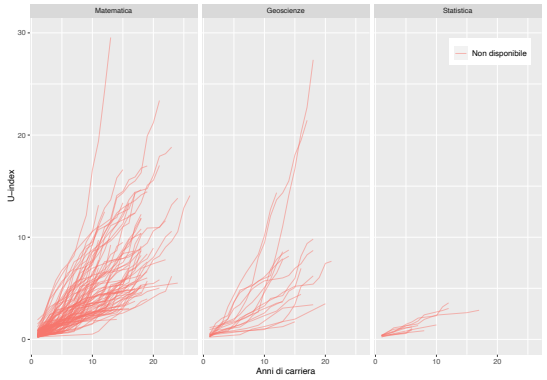

Figura 3.20: U-index

I colori differenti nelle Figure 3.17, 3.18 distinguono il ruolo di Ricercatore e Associato; i segmenti che non hanno inizio in corrispondenza del primo 'Anno di carriera' corrispondono ai docenti che al 31/12/2000 occupavano già la posizione di Professore

associato. Ogni segmento rappresenta l'andamento dell'indice nel tempo, dall'inizio di carriera al raggiungimento della cattedra. Si può osservare che gli andamenti più salienti si hanno negli anni da Ricercatori. L'indice h tende a crescere molto durante questo stadio e quasi ad appiattirsi durante gli anni da Associato; diversamente si comporta u, che continua a crescere per tutta la durata di carriera. Gli indici h e u per i docenti del secondo gruppo sembrano assumere un comportamento lievemente differente. In questo caso per l'area Matematica, dai valori dell'*h-index* si potrebbero riconoscere gli anni in cui un docente da Ricercatore diventa Associato, soltanto guardando le fasi di crescita e successivo stallo dell'indice. Per la categoria di Geoscienze e Statistica, entrambi u e h continuano invece a crescere fino all'anno corrispondente al raggiungimento della massima carica. In entrambi i casi, considerando l'evento di interesse avvenuto in corrispondenza di durate di carriera note con certezza e non, i comportamenti degli indici sembrano essere più interessanti nei 10 anni dalla prima pubblicazione, dove subiscono un'evoluzione maggiore.

3.2 Ricercatori al 2005

Si consideri il secondo dataset costruito nel Capitolo 2, a partire dal campione selezionato di Ricercatori al 31/12/2005. Le informazioni disponibili sono le seguenti:

- $i = 1, ..., 936$ Ricercatori;
- $k = 1, ..., 14\,224$ pubblicazioni dei Ricercatori in esame.

Tra i Ricercatori selezionati, al 31/12/2021 il 17% raggiunge il ruolo di Professore ordinario, il 49% ricopre il ruolo di Associato, il 34% di Ricercatore.

3.2.1 Analisi individuali

Si considerino due docenti dal campione con le medesime durate di carriera: il primo raggiunge il ruolo di Professore ordinario entro il 31/12/2021, il secondo risulta ancora Ricercatore a fine periodo osservazionale. In Figura 3.21, 3.22, 3.23, 3.24 si riportano l'indice h, u, ca e m rispettivamente, per il primo docente considerato:

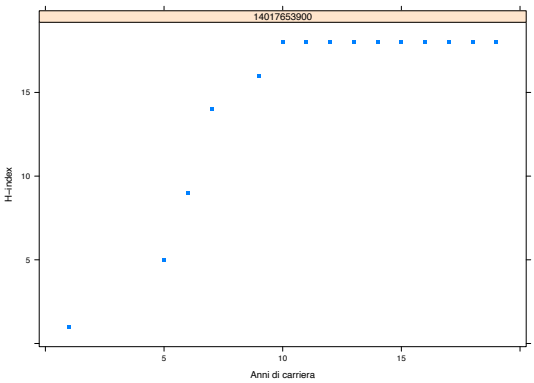

Figura 3.21: H-index

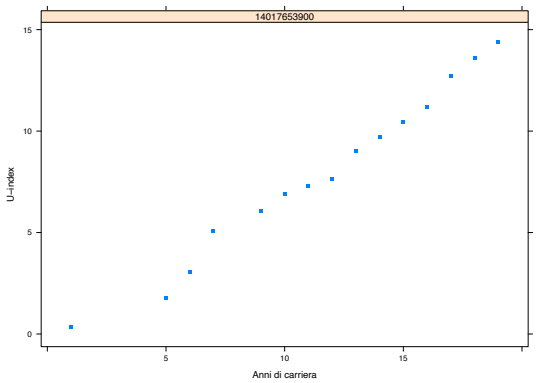

Figura 3.22: U-index

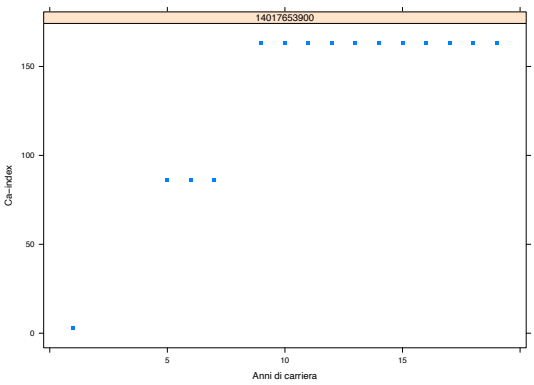

Figura 3.23: Ca-index

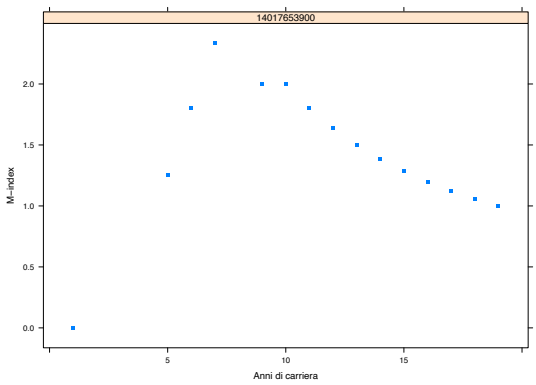

Figura 3.24: M-index

Si osserva che il docente in esame raggiunge il ruolo di Professore ordinario al 19° anno di carriera, con un valore dell'*h-index* pari a 18 e u pari a 14 circa. Il valore di h cresce molto velocemente nei primi 10 anni di carriera, stabilendosi sul valore finale con il quale il docente diventa Professore ordinario. Diversamente, u continua a crescere nel tempo dal 10° anno di carriera in maniera quasi lineare. Confrontando tali andamenti con il secondo docente, in Figura 3.25, 3.26, 3.27, 3.28 si riportano l'indice h, u, ca e m di quest'ultimo:

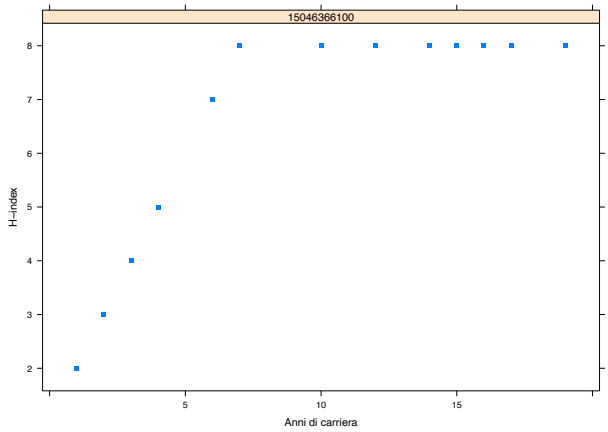

Figura 3.25: H-index

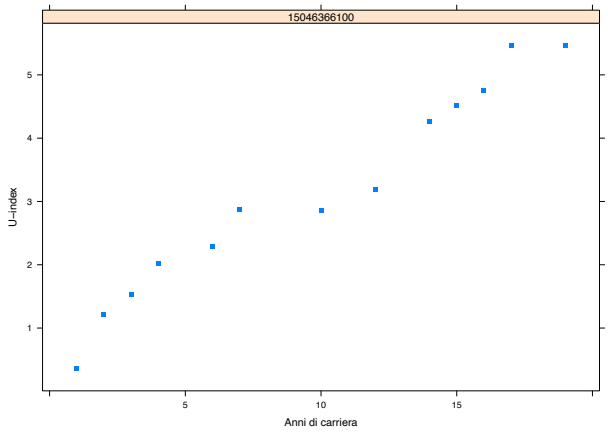

Figura 3.26: U-index

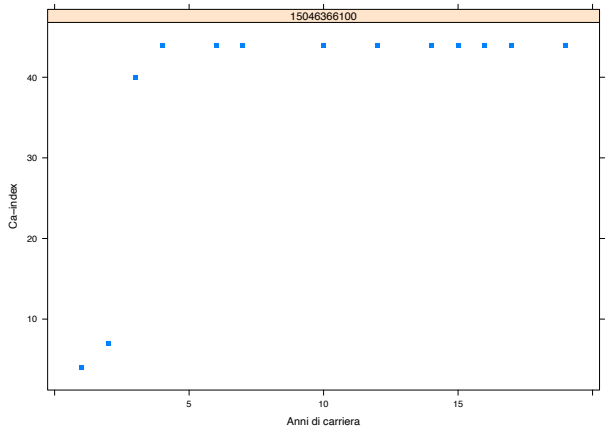

Figura 3.27: Ca-index

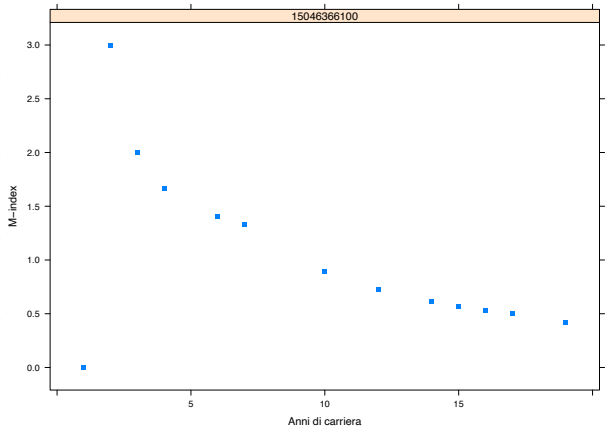

Figura 3.28: M-index

L'indice h per i due docenti è lo stesso al quinto anno di carriera, tuttavia il numero di pubblicazioni effettuate è diverso: avendo considerato una pubblicazione per anno, il primo raggiunge tale valore risultando operativo bibliograficamente al secondo e quinto anno di carriera; il secondo, con almeno una pubblicazione per ogni anno di carriera, dal primo al quinto. Ciò non implica necessariamente che il primo docente abbia pubblicato

di meno rispetto al secondo, ma le sue pubblicazioni hanno avuto un impatto maggiore dal punto di vista citazionale. Notevoli differenze, dopo i 5 anni di carriera, si colgono osservando l'indice u: mentre l'indice per il primo Ricercatore aumenta da 2 a 7 circa nei successivi 5 anni, per il secondo tale valore incrementa da 2 a 3. La crescita più o meno rapida degli indici sembra avere un impatto sulle lunghezze di carriera. Allo stesso modo, la rapidità con cui il *ca-index* aumenta e l'*m-index* diminuisce, sottolineano tale discrepanza.

3.2.2 Analisi collettive

Non per tutti i Ricercatori in esame il 2005 è l'anno di inizio carriera. In Figura 3.29 e 3.30 si osservano i valori degli indici in corrispondenza dell'evento o dell'ultima osservazione disponibile per ognuno, tenendo conto delle durate di carriera:

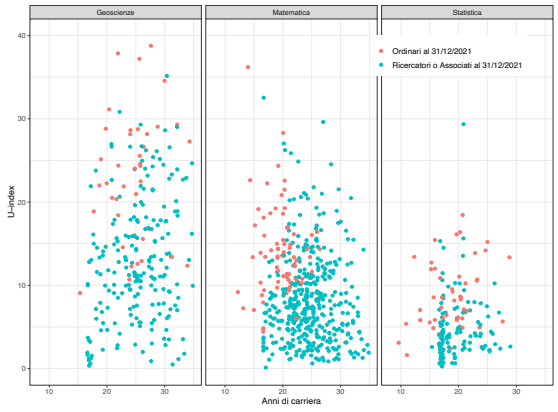

Figura 3.29: Valori dell'indice u in corrispondenza dell'evento/censura

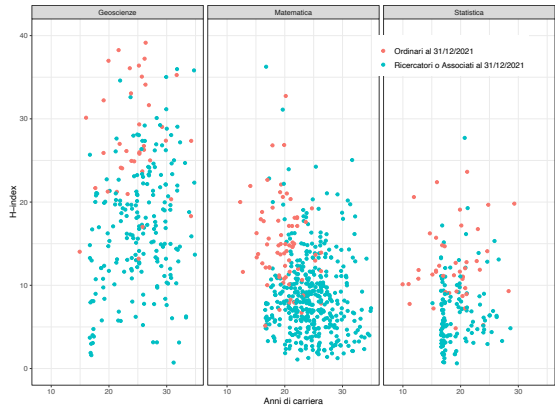

Figura 3.30: Valori dell'indice h in corrispondenza dell'evento/censura

Per entrambi le Figure 3.29 e 3.30 si riportano i valori dei due indici u e h, rispettivamente, sull'asse delle ordinate, in corrispondenza dell'ultima osservazione disponibile per ogni docente. In rosso sono indicati i valori di coloro che raggiungono il ruolo di Professore ordinario entro il 31/12/2021, in verde il valore dell'indice al 31/12/2021 per coloro che non raggiungono la massima carica accademica. Si nota che in generale i valori per l'indice u coi quali i docenti raggiungono la cattedra sono più elevati per la categoria di Geoscienze, rispetto alle altre due. Ancora una volta, come nel precedente campione di Professori ordinari, si osserva che i valori coi quali i docenti raggiungono il ruolo non sono tuttavia molto elevati. Simili considerazioni possono essere effettuate per l'indice h: questo registra valori meno elevati per le categorie di Matematica e Sta-

tistica, in corrispondenza del verificarsi dell'evento.

Si vogliono osservare come gli indici h e u crescono nelle carriere, distinguendo chi sperimenta l'evento e chi no entro la fine del periodo osservazionale. In Figura 3.31, 3.32, 3.33, 3.34 si riportano le distribuzioni degli indici in termini di: variazione del valore dell'indice tra l'inizio di carriera e l'anno in cui viene sperimentato l'evento o in corrispondenza dell'ultimo valore aggiornato al 31/12/2021, diviso per il numero di anni di carriera.

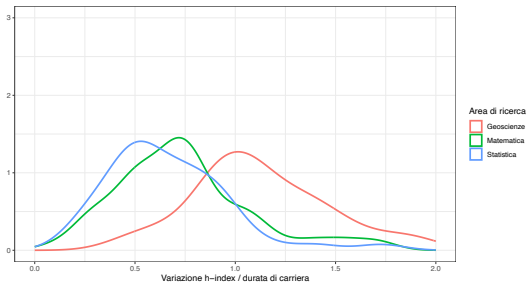

Figura 3.31: Variazione indice h per chi diventa Professore ordinario

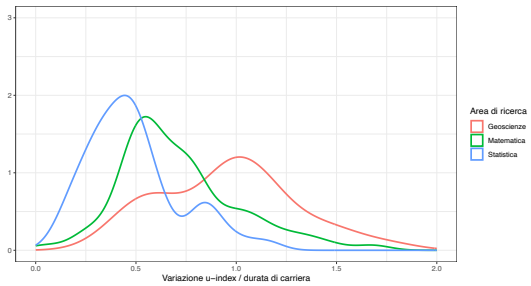

Figura 3.32: Variazione indice u per chi diventa Professore ordinario

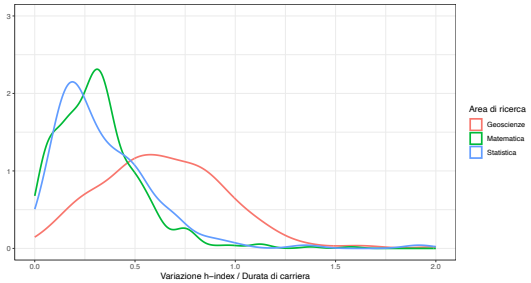

Figura 3.33: Variazione indice h per chi non diventa Professore ordinario

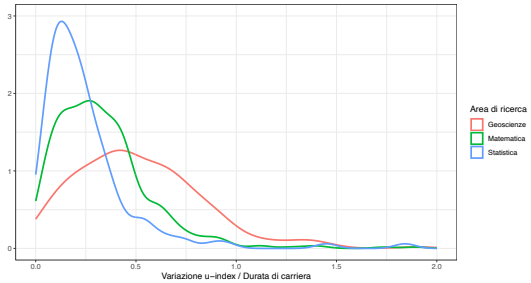

Figura 3.34: Variazione indice u per chi non diventa Professore ordinario

Si può osservare che la distribuzione della variazione dell'indice h, divisa per il numero di anni di carriera è più spostata a sinistra per i Ricercatori che non diventano Professori ordinari entro il 31/12/2021. Le variazione maggiori, anche in relazione al numero di anni di carriera trascorsi, si hanno per la categoria di Geoscienze. Simili osservazioni possono essere effettuate sull'indice u, indicando che nonostante i due indici siano costruiti su scale diverse, le loro variazioni riflettono effetti simili nelle durate di carriera.

3.3 Scelta per i successivi modelli

Si è dunque osservato che l'andamento degli indici nel tempo sembra mostrare alcune caratteristiche comuni tra i docenti. Innanzitutto si è visto che gli indici h e u tendono a crescere più velocemente all'inizio di carriera per i docenti che impiegano un minor numero di anni per diventare Professori ordinari. Inoltre, un valore estremamente elevato, con cui un docente raggiunge la massima carica è osservato raramente.
Considerando i ruoli accademici occupati da ogni docente nel campione di Professori ordinari al 31/12/2021, si è osservato che l'indice h cresce più rapidamente nella fase da Ricercatore piuttosto che negli anni in carica da Professore associato, diversamente dall'indice u che continua a crescere notevolmente. Sembra comunque interessante valutare l'andamento degli indici a diversi stadi di carriera, in particolare nei 10 anni dalla prima pubblicazione, dove si osservano i cambiamenti maggiori.
Tale effetto risulta tuttavia meno interessante per l'indice *ca* e m. Per il *ca-index* si è osservato che l'andamento generale è di crescita: in corrispondenza di alcune annate accademiche, per alcuni docenti, si registrano notevoli aumenti nei valori rispetto a quelli osservati negli anni precedenti. Tale effetto potrebbe essere causato da pubblicazioni di un docente citate da molti autori e per la paternità bibliografica che egli ha su tali pubblicazioni. Un docente con poche pubblicazioni, ma autore di una monografia molto citata avrà un valore dell'indice molto più elevato rispetto a docenti con un numero elevato di pubblicazioni, ma con molti coautori. Infine l'*m-index*, dipendendo dall'indice h, non mostra un effetto particolarmente interessante per le analisi in questione. Si è osservato che diversamente dai tre precedenti, il suo andamento, dopo una prima fase iniziale di crescita, è decrescente: il tasso di crescita è più o meno rapido a seconda del valore assunto da h negli anni.

A questo punto delle analisi si è perciò scelto di proseguire con i soli indici h e u: questi sembrano fornire informazioni più interessanti per quanto concerne lo studio in esame. Avendo osservato che i loro valori risultano più informativi in corrispondenza dei primi 10 anni di carriera, per i due campioni di docenti in esame, si sono creati tre insiemi di osservazioni:

- il primo caratterizzato da tutti i docenti e selezionando le covariate a un anno

dalla prima pubblicazione;

- il secondo caratterizzato dai docenti che raggiungono la massima carica almeno dopo 5 anni dalla prima pubblicazione, e selezionando le covariate relative a cinque anni dalla prima pubblicazione;

- il terzo caratterizzato dai docenti che raggiungono la massima carica almeno dopo 10 anni dalla prima pubblicazione, selezionando le covariate relative a dieci anni dalla prima pubblicazione.

Questi tre insiemi di osservazioni verranno successivamente utilizzati in fase di modellazione: ogni singolo modello verrà stimato utilizzando l'indice u o l'indice h, separatamente, per ogni insieme di osservazioni creato, con i valori delle covariate registrati a uno, cinque, dieci anni di carriera.

Capitolo 4

Analisi di sopravvivenza: previsione degli anni di carriera

La previsione del numero di anni impiegati per raggiungere il ruolo di Professore ordinario è il focus del presente capitolo. Si vuole valutare se e come gli indici bibliometrici siano in grado di fornire utili indicazioni per la previsione degli anni trascorsi fino all'incarico massimo. Per la trattazione di tale argomento si sono utilizzati diversi modelli per l'analisi di durata, utilizzando approcci parametrici, semi-parametrici e non parametrici. L'approccio non parametrico è utilizzato per descrivere i dati rispetto ai fattori di interesse. Tuttavia se si è interessati ad analizzare la relazione tra una serie di covariate e il tempo fino alla realizzazione di un evento, gli approcci semi parametrici e parametrici risultano più appropriati, fornendo stime dell'impatto di ciascun fattore sulla sopravvivenza. Nel seguito del capitolo vengono mostrati l'approccio semi-parametrico attraverso il modello di Cox, con covariate fisse e dipendenti dal tempo, l'approccio parametrico attraverso i modelli GAMM, gli approcci non parametrici di Kaplan-Meier e delle foreste casuali di sopravvivenza.

4.1 Analisi di sopravvivenza

L'analisi di sopravvivenza, o in maniera più generica, l'analisi di durata tratta del tempo fino alla realizzazione di un evento (Altman and Bland, 1998). Nell'analisi di tali dati l'obiettivo di interesse non soltanto è legato al verificarsi o meno di un evento, ma anche all'essere a conoscenza di quando quest'ultimo sia avvenuto. I modelli di sopravviven-

za permettono di includere sia l'informazione relativa alla realizzazione dell'evento che l'aspetto temporale come output del modello e sono strutturati per tener conto di dati censurati.

4.1.1 Modellazione del tempo fino alla realizzazione dell'evento

Nell'analisi dei dati di sopravvivenza è necessario fissare due componenti: la variabile relativa alla durata e l'indicatore dell'evento.
La prima componente è definita attraverso una variabile casuale non negativa, che rappresenta la lunghezza dell'intervallo di tempo che intercorre dall'inizio del periodo osservazionale fino al verificarsi dell'evento o all'ultima osservazione disponibile. Selezionare l'istante di inizio in maniera appropriata è importante al fine di evitare distorsioni. La seconda componente è rappresentata da una variabile categoriale che indica lo stato del soggetto al termine dell'intervallo di tempo considerato. La definizione di questa variabile indicatrice dipende dalla tipologia di dato considerato. Nel presente contesto, le durate non interamente osservate non sono del tutto dati mancanti, ma dati parziali, e identificate con il termine di durate censurate.

Siano, nel presente lavoro, $T_1, ..., T_n$ gli anni impiegati da ciascun docente i, per $i = 1, ..., 759$ per raggiungere il ruolo di Professore ordinario a partire dalla data della pubblicazione del primo articolo. Per ciascun individuo si consideri la coppia di osservazioni (T_i, δ_i), dove T_i rappresenta gli anni di carriera in maniera discreta e δ_i la variabile che indica se la durata è stata osservata ($\delta = 1$) o risulta censurata ($\delta = 0$). Una durata di carriera T_i associata al docente i risulta censurata a sinistra (C_i) se l'evento di interesse, diventare Professore ordinario, è accaduto prima dell'anno 2000: per tali docenti non si conosce con esattezza il momento in cui hanno sperimentato l'evento, ma si sa che l'hanno sperimentato in un certo momento prima di tale anno. Si considera che i soggetti censurati abbiano la stessa probabilità dei soggetti che rimangono nell'analisi, di sperimentare l'evento.

Dal campione di $n = 759$ docenti, 122 raggiungono il ruolo di Professore ordinario prima dell'anno 2000: per tali docenti le durate di carriere sono perciò censurate. Per

tener conto di tale informazione è stata creata una nuova variabile dicotomica, Evento, che vale 1 se il docente ha sperimentato l'evento dopo il 31/12/2000, dunque la durata è nota con certezza, e 0 se censurata.

Essendo il tempo T_i definito in maniera discreta, il rischio istantaneo di sperimentare l'evento all'anno di carriera $j = 1, 2, ...$ può essere espresso come la probabilità condizionata che l'evento si verifichi in un dato istante t_j condizionatamente al fatto che non si sia verificato fino a quel momento (Tutz, Schmid et al., 2016). Il problema dei dati censurati a sinistra può essere trasformato in un problema di dati censurati a destra, moltiplicando ciascun dato per -1 e applicando la teoria classica della censura a destra (Gomez et al., 1992). Si può quindi scrivere: $-Y_i = \min\{-T_i, -C_i\}$. Consderando un individuo con covariate X_i, il rischio di sperimentare l'evento all'anno di carriera t_j si esprime con:

$$\lambda_i(X_i) = \Pr(T_i = t_j | T_i > t_{j-1}, X_i) = \Pr(t_{j-1} < T_i \leq t_j | T_i \geq t_{j-1}, X_i) \tag{4.1}$$

e la funzione di probabilità discreta è data da

$$f_i = \Pr(T_i = t_j | X_i) = S(t_{j-1}|X_i) - S(t_j|X_i) \tag{4.2}$$

La probabilità di sopravvivere oltre un certo istante di tempo t può essere ottenuta come il prodotto delle probabilità di sopravvivenza condizionate per tutti gli istanti di tempo tali per cui $t_j \leq t$. La probabilità di sopravvivenza nei modelli per tempi discreti è data da

$$S_i(t_j|X_i) = \Pr(T_i > t_j | X_i) = \prod_{j:t_j \leq t} (1 - \lambda_i(X_i)) \tag{4.3}$$

Un'assunzione fondamentale al fine di costruire la funzione di verosimiglianza è che le durate e le censure siano tra loro indipendenti. L'osservazione corrispondente a un preciso istante temporale fornisce informazioni sulla probabilità che l'evento si verifichi esattamente in quel momento. Per un'osservazione censurata a sinistra tutto ciò che sappiamo è che l'evento si è già verificato, quindi il contributo alla verosimiglianza è la

distribuzione cumulata valutata all'istante in esame condizionatamente agli istanti di tempo in cui il soggetto è stato osservato, ma nei quali non ha subìto l'evento. La stessa può essere scritta come segue:

$$L(\beta; y) = \prod_{i=1}^{n} f(y_i|x_i)^{\delta_i} F(y_i|x_i)^{1-\delta_i} = \prod_{x \in D} f(y_i|x_i) \prod_{x \in C} F(y_i|x_i) \tag{4.4}$$

con $F(y_i) = 1 - S_i(t_j|X_i)$ funzione di sopravvivenza, D è l'insieme di valori osservati e C il set di valori censurati a sinistra.

4.1.2 Previsione degli anni di carriera

Dalle considerazioni effettuate al termine del Capitolo 3, volendo porre a confronto l'indice h e u, il campione selezionato a partire dai Professori ordinari al 31/12/2021, caratterizzato da 759 docenti e le loro relative $k = 10\,054$ pubblicazioni, è stato suddiviso. Si sono considerati tre diversi insiemi di osservazioni, nel seguente modo.
Per ciascun docente si sono considerate le informazioni contenute nelle variabili 'Anni di carriera', 'Sesso', 'Area di ricerca', 'Anno di inizio carriera', 'Evento', in corrispondenza dell'ultima pubblicazione disponibile per ciascuno. Si sono costruiti tre insiemi di osservazioni, utilizzando le covariate appena citate e estraendo i valori degli indici dal dataset di 10 054 pubblicazioni, in corrispondenza di tre durate di carriera diverse, per ciascuno:

- il primo insieme è caratterizzato da tutti i 759 docenti con valori degli indici a *un anno* da inizio carriera;
- il secondo insieme è caratterizzato da 759 docenti con valori degli indici a *cinque anni* da inizio carriera;
- il terzo insieme è caratterizzato da 714 docenti con valori degli indici a *dieci anni* da inizio carriera, eliminando i docenti che hanno raggiunto il ruolo di ordinario entro i 10 anni dalla prima pubblicazione;

I tre insiemi di osservazioni sono stati utilizzati nella fase di modellazione, dove l'indice h e l'indice u sono stati utilizzati separatamente, assieme alle altre covariate, nella stima dei modelli. In questo modo i modelli ottenuti sono stati confrontati sia in

termini di capacità predittive fornite da ciascuno dei due indici, sia considerando le loro prestazioni a diversi stadi di carriera.

4.2 Stima di Kaplan-Meier

L'approccio più comune in letteratura, adattato ai dati di sopravvivenza, è lo stimatore di Kaplan-Meier (Bland and Altman, 1998). Esso lavora spezzando la stima di *S(t)* in una serie di intervalli basati sugli eventi osservati nel tempo. Le osservazioni in questione contribuiscono alla stima di *S(t)* fino a che l'evento non accade o fino a quando non risultano censurate. Per ogni intervallo di tempo, la probabilità di sopravvivere fino al termine dell'intervallo è calcolata considerando che i soggetti sono a rischio di subire l'evento all'inizio dell'intervallo. La curva per *S(t)* può essere rappresentata come una funzione a gradini con il tempo sull'asse delle ascisse.
La stima di Kaplan-Meier rappresenta una delle migliori opzioni da utilizzare per misurare la frazione di soggetti che sopravvivono fino a un certo istante di tempo, senza subire l'evento. Tre assunzioni devono essere rispettate per poter utilizzare tale tecnica di stima non parametrica:

- l'evento di interesse avviene in un preciso istante temporale;
- la probabilità di sopravvivenza è la stessa per tutti i soggetti che entrano nel campione, indipendentemente dal momento in cui sono entrati nello studio osservazionale;
- i soggetti censurati hanno le stesse prospettive di sopravvivenza (stesso rischio) di quelli che ancora non hanno subìto l'evento.

Nella realtà, la vera funzione di sopravvivenza non si conosce. Tale stimatore approssima la vera funzione di sopravvivenza utilizzando i dati raccolti. Lo stimatore è definito come la frazione dei soggetti sopravvissuti per un certo ammontare di tempo sotto le stesse circostanze, dato dalla seguente formula:

$$\hat{S}(t) = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i}) \tag{4.5}$$

dove t_i indica l'istante in cui almeno un evento si è verificato, d_i il numero di eventi accaduti in t_i e n_i rappresenta il numero di individui sopravvissuti fino a t_i. La proba-

bilità di sopravvivenza al tempo t è uguale al prodotto delle probabilità di sopravvivere fino a quell'istante.

Volendo osservare come i valori degli indici considerati a diversi anni di carriera influiscano sul rischio di diventare Professore ordinario, si considerano i tre insiemi di osservazioni, contenenti le covariate a uno, cinque, dieci anni dalla prima pubblicazione. Si riportano in Tabella 4.1 le distribuzioni degli indici nei tre insiemi:

Indice	**Minimo**	**1 Quartile**	**2 Quartile**	**Media**	**3 Quartile**	**Massimo**
U a 1 anno	0.05	0.33	0.34	0.45	0.49	2.46
H a 1 anno	1.00	1.00	1.00	1.32	1.00	5.00
U a 5 anni	0.06	0.65	1.43	1.40	1.89	6.51
H a 5 anni	1.00	2.00	3.00	3.97	5.00	17.00
U a 10 anni	0.16	2.01	3.33	3.94	5.15	23.50
H a 10 anni	1.00	6.00	9.00	10.18	13.00	40.00

Tabella 4.1: Distribuzioni degli indici nei diversi stadi di carriera

Osservando le distribuzioni nei diversi insiemi a uno, cinque, dieci anni dalla prima pubblicazione, i valori medi degli indici sono più vicini considerando il primo e quinto anno di carriera, rispetto al 10°. In particolare si osserva che per l'indice u il valore medio al 10° anno di carriera è pari a quattro, mentre valori più elevati risultano rari. In maniera analoga l'indice h ha un valore medio pari a 10 in corrispondenza del 10° anno di carriera e più rari sono i valori elevati.

Per stabilire se l'effetto degli indici sul rischio sia lo stesso, condizionatamente ai valori che essi assumono in corrispondenza di specifiche durate di carriera, si sono suddivisi h e u in classi differenti, all'interno di ogni insieme di osservazioni considerato. Poiché i valori degli indici a inizio carriera sono piuttosto bassi, si sono considerate due classi: per u [0, 1], (1,3] e per h, [1,3], (3,5]. Le curve di sopravvivenza ottenute, considerando tutti i docenti nell'insieme, sono riportate nelle Figure 4.1 e 4.2:

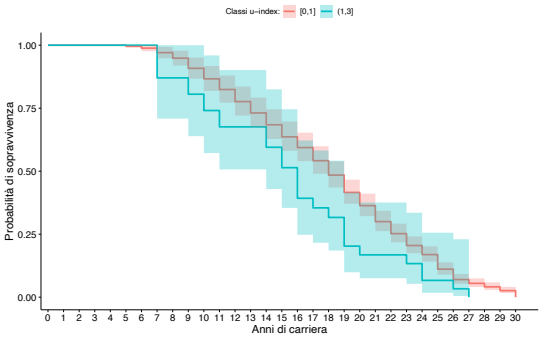

Figura 4.1: KM con valori dell'indice u a 1 anno di carriera

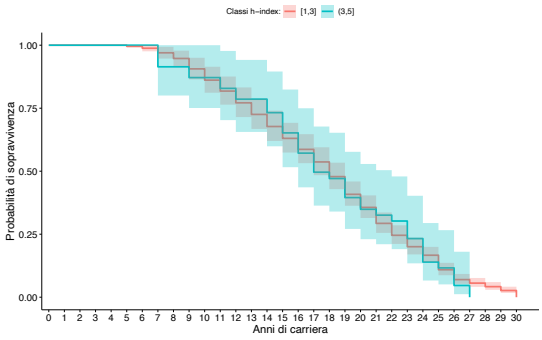

Figura 4.2: KM con valori dell'indice h a 1 anno di carriera

Un valore dell'indice $u > 1$, a inizio carriera, sembra avere un effetto sul rischio di sperimentare l'evento maggiore rispetto a un valore < 1, per qualsiasi durata di carriera. Diversamente, un valore di h maggiore o minore di tre a un anno dalla prima pubblicazione, non mostra un effetto diverso sulla probabilità di sperimentare l'evento in questione.
Considerando i gruppi di osservazioni con gli indici a cinque e dieci anni dall'inizio di carriera, gli intervalli di valori considerati sono i seguenti:

- *u-index*: classi $[0, 3]$, $(3, 7]$ a 5 anni e $[0, 3]$, $(3, 7]$, $(7, 24]$ a 10 anni;
- *h-index*: classi $[1, 3]$, $(3, 10]$, $(10, 17]$ a 5 anni e $[1, 3]$, $(3, 10]$, $(10, 40]$ a 10 anni.

Poiché i valori dell'indice u a cinque anni di carriera sono molto vicini a quelli considerati a un anno dalla prima pubblicazione, si sono considerate solo due classi. Per gli indici a 10 anni, si sono considerate le stesse classi di valori del 5° anno dalla prima pubblicazione, estendendo gli intervalli nei valori massimi.

Le curve di sopravvivenza sono riportate nelle Figure 4.3, 4.4, 4.5, 4.6:

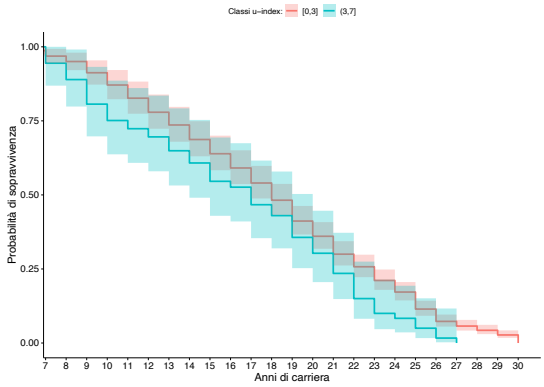

Figura 4.3: Con valori dell'indice u a 5 anni di carriera

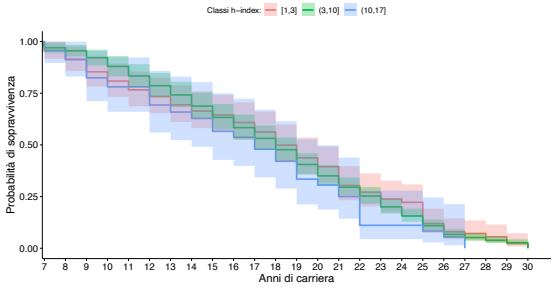

Figura 4.4: KM con valori dell'indice h a 5 anni di carriera

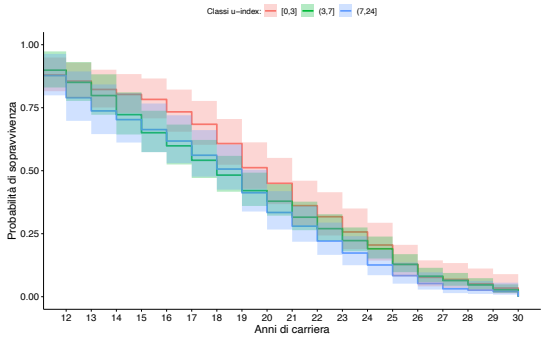

Figura 4.5: KM con valori dell'indice u a 10 anni di carriera

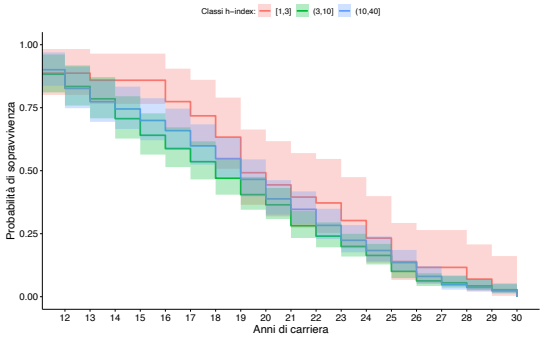

Figura 4.6: KM con valori dell'indice h a 10 anni di carriera

Si nota chiaramente che i docenti con valori dell'indice u più basso, a 5 e 10 anni di carriera, hanno una probabilità maggiore di sopravvivenza, ossia un rischio minore di sperimentare l'evento, per qualsiasi durata di carriera. Tale effetto è più forte considerando l'indice a 10 anni dalla prima pubblicazione. Osservando come i valori di u agiscono sulla probabilità di sperimentare l'evento a 10 anni da inizio carriera, si osserva che il rischio di sperimentare l'evento è circa lo stesso per i docenti con un indice u compreso nella classe $(3, 7]$ o $(7, 24]$. Per l'indice h, considerando i valori a 5 anni dalla prima pubblicazione, l'effetto sul rischio di un valore più o meno grande è circa lo stesso considerato qualsiasi durata di carriera. Un docente con indice $h < 3$, a 10 anni dalla prima pubblicazione, ha un rischio nettamente minore di sperimentare l'evento, rispetto a un docente con valore dell'indice più elevato. Inoltre un docente con valore dell'indice compreso tra 3 e 10, se si considerano durate di carriera superiori ai 14 anni di durata, ha un rischio più elevato di sperimentare l'evento rispetto a un docente con valore di h più elevato.

4.3 Modello Proporzionale di Cox

Il modello di Cox a rischi proporzionali (Cox, 1972) è il modello statistico più utilizzato per l'analisi di sopravvivenza. È un modello di regressione sul tempo che descrive la relazione tra l'incidenza di un evento, espresso dalla funzione di rischio, e un insieme di covariate. Tale modello si basa sull'assunzione che la funzione di rischio, dato un insieme di variabili esplicative $X_1, ..., X_p$, sia proporzionale a una funzione di rischio di base definita per l'individuo i:

$$\lambda(t_i|X_i) = \lambda_0(t) \cdot e^{X_i\beta} \tag{4.6}$$

dove $\beta = (\beta_1, \beta_2, ..., \beta_p)$ è un vettore di coefficienti di regressione, $\lambda_0(t)$ è il rischio di base (cioè la probabilità che si verifichi l'evento, dato che non si è verificato prima, quando tutte le covariate sono uguali a zero) e X_i il vettore di covariate fissate nel tempo.
Si tratta di un approccio semi parametrico in quanto il modello contiene una componente non parametrica e una componente parametrica. Il rischio di base $\lambda_0(t)$ è stimato in via non parametrica, quindi diversamente dalla maggior parte dei modelli statistici, i tempi di sopravvivenza non seguono una particolare distribuzione statistica e la forma del rischio è arbitraria. La funzione del rischio di base non ha bisogno di essere stimata per fare inferenza sul rischio relativo: poiché non è vulnerabile a errate specificazioni del rischio di base il modello di Cox è più robusto rispetto agli approcci parametrici.
La componente parametrica è insita nel vettore di covariate. Quest'ultimo moltiplica il rischio di base per lo stesso ammontare indipendentemente dall'istante di tempo considerato, così che l'effetto di qualsiasi covariata sia lo stesso durante l'intero periodo di osservazione.

4.3.1 Stima

Non disponendo di un modello generatore dei dati, e di conseguenza di uno specifico metodo di confronto, il modello di Cox è stato adattato scegliendo due diverse strade. Disponendo del campione di 759 Professori ordinari al 31/12/2021:

- 1° approccio: il 75% delle osservazioni è stato utilizzato per stimare il modello, il 25% per convalidarlo tramite la log-verosimiglianza predittiva;
- 2° approccio: tutti i dati sono stati utilizzati per la stima del modello e il confronto è avvenuto utilizzando i criteri di selezione automatica.

Il modello di Cox è stato inoltre adattato scegliendo due diversi approcci:

- il modello di Cox a rischi proporzionali;
- il modello lineare penalizzato di Cox a rischi proporzionali (Simon et al., 2011).

In Tabella 4.2 le stime dei coefficienti relativi agli indici ottenuti:

Metodo	**Modello con indice**	**Stima e Verifica**					**Tutti i dati**				
		$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	z	$\Pr(>\|z\|)$	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	z	$\Pr(>\|z\|)$
Cox	U a 1 anno	0.362	1.436	0.185	1.953	0.051	0.311	1.365	0.166	1.876	0.061
Cox penalizzato	U a 1 anno	0.305	-	-	-	-	0.274	-	-	-	-
Cox	H a 1 anno	0.0165	1.0167	0.065	0.252	0.801	0.039	1.039	0.058	0.668	0.504
Cox penalizzato	H a 1 anno	0.018	-	-	-	-	0.016	-	-	-	-
Cox	U a 5 anni	0.03	1.03	0.05	0.601	0.548	0.0549	1.056	0.042	1.308	0.191
Cox penalizzato	U a 5 anni	0.023	-	-	-	-	0.044	-	-	-	-
Cox	H a 5 anni	-0.007	0.993	0.018	-0.361	0.718	0.005	1.005	0.015	0.317	0.751
Cox penalizzato	H a 5 anni	-0.007	-	-	-	-	0.002	-	-	-	-
Cox	U a 10 anni	0.040	1.041	0.02	2.064	0.039 *	0.032	1.032	0.017	1.907	0.057
Cox penalizzato	U a 10 anni	0.0343	-	-	-	-	0.027	-	-	-	-
Cox	H a 10 anni	0.008	1.008	0.008	1.000	0.317	0.004	1.004	0.007	0.531	0.596
Cox penalizzato	H a 10 anni	0.006	-	-	-	-	0.003	-	-	-	-

Tabella 4.2: Stime dei coefficienti

Da sinistra a destra in Tabella 4.2 sono riportate le stime dei coefficienti relativi agli indici stimati nei modelli, rispettivamente su una parte delle osservazioni e utilizzando tutti i dati. I coefficienti β ottenuti dall'adattamento del modello di Cox possono essere interpretati in termini di rischio attraverso la trasformazione esponenziale $\exp(\beta)$. Tale modello fornisce inoltre una stima dell'errore standard, la statistica test e il livello di significatività osservato.
Il coefficiente relativo all'indice h, al netto delle altre variabili, non è significativo a nessuno stadio di carriera considerato; l'indice u mostra un coefficiente significativo al 5%, soltanto per i valori considerati a 10 anni da inizio carriera. Essendo il coefficiente β per l'indice u maggiore di zero, un valore elevato dell'indice incrementa il rischio di divenire Professore ordinario, rispetto a un valore piccolo. Tuttavia, considerando i valori dell'indice a uno, cinque, dieci anni dalla prima pubblicazione, i coefficienti hanno un effetto sul rischio sempre minore, il che implica che a parità del valore dell'indice

considerato, ad ogni stadio, è necessario un numero di anni di carriera maggiore per sperimentare l'evento. Il coefficiente β relativo all'indice h è molto vicino allo zero a un anno dalla prima pubblicazione, e negativo a 5 anni da inizio carriera. Questo implica che un docente con valore dell'indice h elevato, a 5 anni di carriera, non ha un rischio maggiore di sperimentare l'evento rispetto a un docente con un valore piccolo dell'indice. Il modello di Cox penalizzato fornisce invece solo una stima dei coefficienti β: queste non risultano molto diverse da quelle del modello di Cox semplice.
Si può osservare che in entrambi i casi, utilizzando una porzione di dati o tutte le osservazioni disponibili, le stime dei coefficienti sono molto simili, così come gli errori standard.

4.3.2 Previsione

Log-verosimiglianza predittiva

Un modo per valutare la stima dei modelli, ed effettuarne un confronto, è attraverso il logaritmo della densità predittiva, anche chiamata log-verosimiglianza predittiva. Il logaritmo della densità predittiva ha un ruolo importante nel confronto dei modelli statistici in particolare per le sue connessioni con la divergenza di informazione di Kullback-Leibler (Kullback and Leibler, 1951).
Non disponendo di un modello generatore dei dati, in questo caso, un modo per stimare la log-verosimiglianza predittiva è considerando la stessa nella stima di massima verosimiglianza (Bodapati and Gupta, 2004). Dopo aver adattato il modello alla porzione di dati utilizzata per la stima, l'accuratezza predittiva viene stabilita sull'insieme di verifica. La verosimiglianza di sopravvivenza per l'individuo i è definita come:

$$L = \prod_{i=1}^{n} [\Pr(T_i = t_{j_i})]^{\delta_i} [\Pr(T_i > t_{j_i})]^{1-\delta_i} = \prod_{i=1}^{n} \prod_{j=1}^{j_i} \lambda_{ij}(X_i)^{\delta_i} (1 - \lambda_{ij}(X_i)^{1-\delta_i} \qquad (4.7)$$

dove t_{j_i} è il tempo di sopravvivenza per il soggetto i, $\lambda_{ij}(X_i)$ la probabilità specifica del soggetto i e X_i le covariate fisse nel tempo.

Si assume l'esistenza di un gruppo di individui, $R(t_j)$, a rischio di sperimentare l'evento al tempo t_j. La probabilità condizionata che il singolo evento accada all'individuo

i, dato l'insieme di covariate X_i associate, è rappresentata da:

$$\frac{\lambda(t_{(i)}|x_{(i)})}{\sum_{i\in R(t_j)}\lambda(t_{(i)}|x_{(i)})} = \frac{\lambda_0(t)e^{X_i\beta}}{\sum_{i\in R(t_j)}\lambda_0(t)e^{X_i\beta}} = \frac{e^{X_i\beta}}{\sum_{i\in R(t_j)}e^{X_i\beta}} \tag{4.8}$$

L'equazione è data dal rapporto tra la funzione di rischio per l'individuo i in un tempo t_j, e la somma delle funzioni di rischio per tutti gli individui in $R(t_j)$. Tale formula è definita funzione di verosimiglianza parziale in quanto il modello di Cox considera la probabilità soltanto per i soggetti che sperimentano l'evento in studio.

Dati i coefficienti $\hat{\beta}$ stimati dai modelli e utilizzando le covariate dell'insieme di verifica, la misura dell'accuratezza predittiva di $\hat{\beta}$ è la log-verosimiglianza predittiva (PLL). La PLL nel presente contesto è definita come segue:

$$PLL(\hat{\beta}) = \log\left(\prod_{i=1}^{m}\left[\frac{e^{X_{iV}\hat{\beta}_S}}{\sum_{j\in R(t_i)}e^{X_{jV}\hat{\beta}_S}}\right]\right) \tag{4.9}$$

dove $i = 1, ..., m$ rappresentano i tempi distinti in cui si verifica l'evento in studio ($n-m$ eventi censurati), con S si indica la provenienza dei $\hat{\beta}$ dall'insieme di stima e con V che le covariate provengono dall'insieme di verifica. Si noti che il numeratore della verosimiglianza dipende solo dalle informazioni degli individui che hanno già sperimentato l'evento, mentre il denominatore utilizza le informazioni su tutti gli individui che non hanno ancora sperimentato l'evento.
Tale metrica consente di misurare le prestazioni predittive di ciascun modello stimato in un campione che non utilizza gli stessi dati della stima. In maniera analoga alla minimizzazione di un criterio di divergenza, il modello migliore è quello identificato massimizzando tale quantità.

Criteri di selezione automatica

Un approccio alternativo per valutare l'accuratezza predittiva dei modelli si basa su misure probabilistiche, che mirano a quantificare sia le prestazioni del modello, obiettivo di interesse in questo caso, che la complessità dello stesso. Si sono utilizzati il criterio di informazione di Akaike (AIC) e il criterio di Schwarz (BIC) per il confronto in termini predittivi, stimando i modelli sull'intero campione di docenti a disposizione. In maniera simile alla massimizzazione della log-verosimiglianza predittiva, dato un insieme di

modelli, il migliore risulta essere quello che minimizza rispettivamente le quantità

$$AIC = -2\log(L) + 2p \tag{4.10}$$

e

$$BIC = -2\log(L) + p\log(n) \tag{4.11}$$

dove $\log(L)$ è la log-verosimiglianza stimata su tutti i dati, p il numero di covariate inserite nel modello e n il numero di docenti.

4.3.3 Risultati

Si riportano in Tabella 4.3 le metriche di accuratezza dei modelli stimati:

Confronto	**Metodo**	**Modello con indice**					
		U a 1 anno	**H a 1 anno**	**U a 5 anni**	**H a 5 anni**	**U a 10 anni**	**H a 10 anni**
PLL	Veros. parziale	-685.76	-685.49	-630.2	-629.82	-585.29	-585.39
	Veros. parziale penalizzata	-346.28	-346.37	-323.88	-323.54	-290.21	-290.29
AIC	Veros. parziale	6952	6955	6943	6945	6487	6491
	Veros. parziale penalizzata	2806	2809	2798	2799	2520	2523
BIC	Veros. parziale	6975	6978	6967	6968	6510	6513
	Veros. parziale penalizzata	2829	2832	2821	2822	2543	2546

Tabella 4.3: Confronto dei modelli stimati

Si può osservare che per tutti i modelli stimati con u o h, utilizzando il metodo della verosimiglianza parziale semplice o penalizzata, la log-verosimiglianza predittiva dà risultati molto simili. Confrontando i modelli in esame attraverso i criteri di selezione automatica, si notano migliori capacità predittive del modello stimato con l'indice u rispetto ad h utilizzando le covariate, in tutti gli stadi di carriera, essendo il valore di AIC e BIC inferiore in corrispondenza dei modelli stimati con u. L'effetto maggiore nell'utilizzare u o h si nota in particolare considerando i modelli contenenti i valori degli indici a uno e dieci anni dalla prima pubblicazione. Si osserva inoltre che utilizzare le covariate a 10 anni di carriera, rispetto ai valori negli anni precedenti, fa sì che sia u che h siano più predittivi sulla risposta.

La log-verosimiglianza predittiva definita in Equazione 4.9 può essere riscritta nel seguente modo:

$$PLL(\hat{\beta}) = \sum_{i=1}^{m} \left(\log \left(e^{X_{iV}\hat{\beta}_S} \right) - \log \left(\sum_{j \in R(t_i)} e^{X_{jV}\hat{\beta}_S} \right) \right) \tag{4.12}$$

con $i = 1, ..., m$ tempi distinti in cui l'evento in studio si verifica. La PLL per ogni modello stimato è dunque calcolata come somma di singole quantità, in corrispondenza di ciascun tempo in cui l'evento si verifica. Tali quantità sono le differenze tra logaritmi definite nell'Equazione 4.12. La PLL è stata calcolata, separatamente, per il modello stimato con l'indice u o con l'indice h.

In ordinata nelle Figure 4.7, 4.8, 4.9 sono riportate le differenze di PLL tra u e h, in ascissa i valori degli indici corrispondenti a ciascun tempo in cui si verifica l'evento:

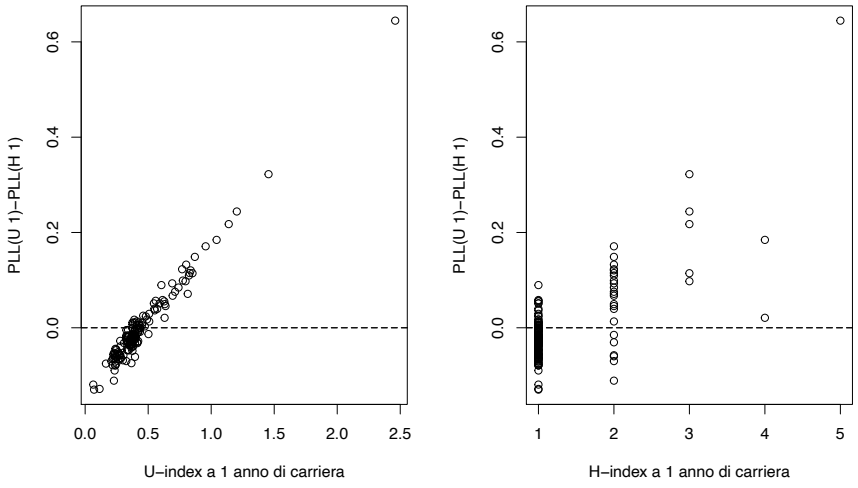

Figura 4.7: Differenza PLL tra l'indice u e h a un anno dalla prima pubblicazione rispetto ai valori dell'indice u a sinistra, h a destra

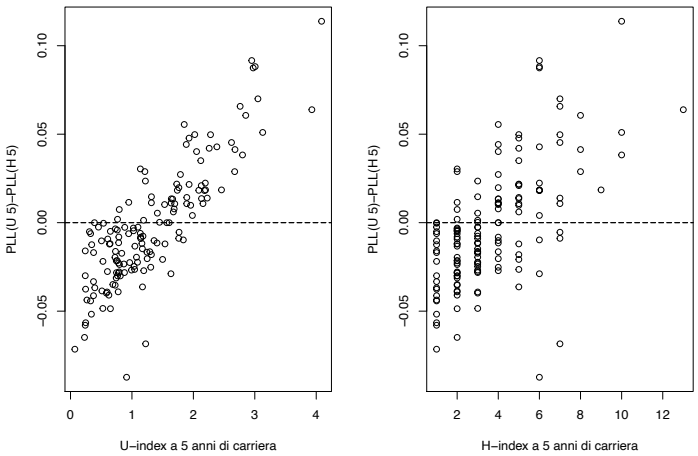

Figura 4.8: Differenza PLL tra l'indice u e h a cinque anni dalla prima pubblicazione rispetto ai valori dell'indice u a sinistra, h a destra

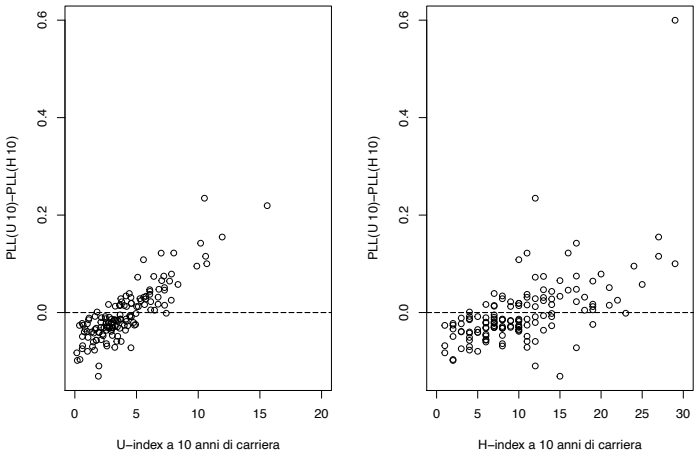

Figura 4.9: Differenza PLL tra l'indice u e h a dieci anni dalla prima pubblicazione rispetto ai valori dell'indice u a sinistra, h a destra

Avendo osservato in Tabella 4.3 che le PLL sono quantità negative e che il modello migliore risulta essere quello con log-verosimiglianza predittiva negativa più bassa, osservazioni corrispondenti a $PLL_{U-index} - PLL_{H-index} > 0$ indicano una migliore capacità predittiva del modello contenente l'indice u rispetto a quello contenente l'indice h. Aver rappresentato le differenze tra PLL in funzione di u e h, permette di osservare i valori degli indici in corrispondenza del verificarsi dell'evento. Si nota che sono i valori compresi rispettivamente tra 1 e 5, e 1 e 15, quelli con i quali i docenti raggiungono il

ruolo di Professore ordinario.

4.4 Modello di Cox: covariate dipendenti dal tempo

Il modello di Cox permette anche di includere covariate che variano nel tempo. In questo caso *u-* e *h-index* dipendono dal tempo e possono opportunamente essere inserite nel modello. Il rischio che il docente *i-esimo*, con vettore di covariate che dipende dal tempo $X_i(t)$, sperimenti l'evento di interesse al tempo t può essere definito nel seguente modo:

$$\lambda_i(t|X_i(t)) = \lambda_0(t) \exp(\beta_1^T x_{i1}(t) + ... + \beta_p^T x_{ip}(t)) \tag{4.13}$$

dove $\lambda_i(t|X_i(t))$ è la funzione di rischio, $\lambda_0(t) \geq 0$ è il rischio di base, $\exp(\beta_1^T x_{i1}(t) + ... + \beta_p^T x_{ip}(t)) > 0$ indica l'effetto delle covariate sul rischio di base ed è la parte parametrica del modello. Solo le covariate dipendono dal tempo in questo caso, mentre i coefficienti β_j per $j = 1, ..., p$ sono costanti al variare di t.

Al fine di ottenere le quantità necessarie per la stima del modello, si è considerato il dataset contenente le pubblicazioni per ogni docente, caratterizzato da $n = 10\,054$ pubblicazioni. Al fine di stimare il modello di Cox con covariate dipendenti dal tempo, il dataset è stato così modificato:

1. per ciascun docente vengono costituiti i sotto-episodi contigui in base al tempo di cambiamento di stato delle covariate h e u, utilizzando periodi di ampiezza pari a un anno;

2. ogni riga del dataset (periodo) rappresenta un intervallo di tempo aperto a sinistra e chiuso a destra;

3. la variabile 'Evento' è stata ricodificata: vale 1 se l'intervallo termina con il verificarsi dell'evento e la durata è nota con esattezza, 0 se la durata non è nota con esattezza e dunque l'intervallo è censurato;

4. per ciascun docente il numero di intervalli è variabile in base all'ultimo sotto-episodio.

Il cambiamento di stato o delle covariate accaduto nell'anno di carriera x entrerà nella funzione di rischio a partire dall'intervallo $(x, x + 1]$.

Si consideri, ad esempio, uno dei docenti tra quelli in esame, il quale raggiunge il ruolo di Professore ordinario al 15° anno di carriera. Considerando il cambiamento di stato delle due covariate, h e u, per il docente in questione, si ottengono i sotto-episodi contenuti in Tabella 4.4:

	Anni di carriera	Scopus ID	Sesso	U-index	H-index	tstart	tstop	Evento	Anno di inizio carriera	Area di ricerca
1	15	13410550200	M	0.4	1	0	1	0	1992	Matematica
2	15	13410550200	M	0.81	2	1	2	0	1992	Matematica
3	15	13410550200	M	1.55	4	2	4	0	1992	Matematica
4	15	13410550200	M	1.89	5	4	7	0	1992	Matematica
5	15	13410550200	M	2.14	6	7	9	0	1992	Matematica
6	15	13410550200	M	2.47	7	9	10	0	1992	Matematica
7	15	13410550200	M	3.08	7	10	12	0	1992	Matematica
8	15	13410550200	M	3.54	7	12	14	0	1992	Matematica
9	15	13410550200	M	3.87	7	14	15	1	1992	Matematica

Tabella 4.4: Alcune righe del dataset selezionato a partire dal campione di Professori ordinari al 31/12/2021

Ogni periodo è stato creato utilizzando il cambiamento di stato delle due covariate. Si osservi che se soltanto uno tra i due valori varia nel tempo, quello fisso viene ripetuto fino a quando non viene raggiunto l'intervallo in cui cambia di stato. Nel caso in esame, l'indice h dall'intervallo $(9, 10]$ fino al termine del periodo osservazionale (intervallo $(14, 15]$) non varia, e viene perciò ripetuto.

4.4.1 Stima del modello

Si sono quindi stimati, considerando le pubblicazioni $k = 1, ..., 10\,054$, i due modelli, contenenti il primo l'indice u, il secondo l'indice h, assieme alle altre covariate. In Tabella 4.5 si riportano le stime dei coefficienti ottenute dai due modelli:

Modello	**Coefficiente**	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	z	$\Pr(>\|z\|)$
Cox con indice u	Sesso: M	0.363	1.437	0.091	3.991	6.57e-05 ***
	Area: Geoscienze	-0.493	0.612	0.108	-4.577	4.71e-06 ***
	Area: Statistica	0.415	1.514	0.105	3.946	7.94e-05 ***
	U-index	0.008	1.008	0.006	1.311	0.19
	Anno di inizio carriera	0.056	1.057	0.008	7.226	4.97e-13 ***
Cox con indice h	Sesso: M	0.378	1.458	0.091	4.160	3.19e-05 ***
	Area: Geoscienze	-0.447	0.64	0.12	-3.731	0.000191 ***
	Area: Statistica	0.39	1.478	0.104	3.764	0.0002 ***
	H-index	0.0002	1.0002	0.006	0.035	0.972
	Anno di inizio carriera	0.058	1.06	0.007	7.841	4.46e-15 ***

Tabella 4.5: Stime dei modelli con covariate dipendenti dal tempo

Dai risultati si osserva che tutte le covariate che non dipendono dal tempo hanno un effetto significativo sul rischio di sperimentare l'evento. In particolare i docenti di Statistica, al netto delle altre covariate hanno un rischio più elevato rispetto ai docenti

di Matematica di diventare Professori ordinari; al contrario, il coefficiente relativo ai docenti di Geoscienze, mostra che al netto delle altre covariate, appartenere a tale categoria il rischio di diventare Professore ordinario è inferiore rispetto a un docente di Matematica. Osservando la covariata relativa all'anno di inizio carriera, da entrambi i modelli, è evidente che c'è un rischio di sperimentare l'evento leggermente più elevato per i docenti che hanno cominciato le carriere in tempi più recenti. Le covariate dipendenti dal tempo non risultano significative: entrambi i coefficienti relativi agli indici non mostrano un effetto significativo sul rischio. In particolare l'indice h non mostra alcun effetto sul rischio di sperimentare l'evento; l'indice u, nonostante non abbia un coefficiente significativo, mostra un effetto sul rischio leggermente superiore rispetto ad h. In entrambi i casi sembra che la dipendenza dal tempo dei due indici non abbia influenza sul rischio di sperimentare l'evento.

4.5 Modello GAMM: covariate dipendenti dal tempo

Si vuole osservare se utilizzando una modellazione diversa per i coefficienti che dipendono dal tempo, questi abbiano un effetto sul rischio di sperimentare l'evento.
Un'ampia classe di modelli per i dati di sopravvivenza può essere rappresentata attraverso modelli misti additivi generalizzati (GAMM, Lin and Zhang (1999)). Con tale rappresentazione, che richiede una specifica trasformazione dei dati originali, è possibile includere nella stima dei modelli effetti non-lineari, spaziali e casuali che variano nel tempo.
Tale modellazione nasce dallo sviluppo dei modelli esponenziali a tratti (PEM, *Piecewise Exponential Model*), che sono essenzialmente modelli lineari generalizzati di Poisson con verosimiglianze proporzionali alla verosimiglianza (parziale) di un modello di Cox (Friedman, 1982). La rappresentazione PEM richiede una suddivisione degli episodi per ciascun soggetto in un numero finito di intervalli e assume che i tassi di rischio siano costanti in ciascuno di essi. Un modello di Cox può essere stimato attraverso un modello di regressione di Poisson, specificando un parametro per ogni intervallo temporale. I risultati sono analoghi anche in termini di errori standard, in quanto la verosimiglianza massimizzata è la stessa.
La scelta degli intervalli di tempo da utilizzare per suddividere gli episodi ha sempre rappresentato una forte criticità per i modelli PEM: se il numero selezionato è molto

piccolo la funzione di rischio potrebbe risultare troppo approssimata. Dall'altro lato, un numero troppo elevato di intervalli potrebbe portare a stime instabili, visto che la funzione di rischio richiede di essere ricalcolata per ciascuno di questi. Con l'introduzione dei modelli additivi tale problema è stato risolto permettendo al rischio e agli effetti variabili nel tempo di entrare nel modello in maniera semiparametrica: tale estensione dei modelli PEM è definita PAMM (*Piece-wise Exponential Additive Mixed Model*, Bender, Groll and Scheipl (2018)). La differenza sostanziale tra i due si ha nel diverso approccio per la stima del rischio di base e per l'introduzione di lisciatori nel modello. Per un generico modello PAMM, il tasso di rischio stimato al tempo t per l'individuo i con vettore di covariate X_i è dato da

$$\lambda_i(t|X_i) = \exp\left(f_0(t_j) + \sum_{k=1}^{p} f_k(X_{i,k}, t_j) + b_{l_i}\right), \quad \forall t \in (k_{j-1}, k_j], \tag{4.14}$$

dove $f_0(t_j)$ rappresenta il logaritmo del rischio di base e $f_k(X_{i,k}, t_j)$, $k = 1,...,p$ denota le tipologie di effetti, di uguale o diversa complessità e potenzialmente dipendenti sia da una covariata che dal tempo. In particolare, sono inclusi lisciatori non-lineari e effetti di lisciamento con effetto variabile nel tempo delle covariate $X_{\cdot k} = (X_{1,k}, ..., X_{n,k})^T$. La variabile t_j è costante in ciascun intervallo per assicurare che il tasso di rischio stimato corrisponda ancora ad un modello PEM. Infine b_{l_i} denota eventuali termini a intercetta casuale inclusi dove l_i, $l = 1, ..., L$ è il gruppo a cui il soggetto appartiene. Quest'ultimo termine non verrà trattato nel presente caso (si farà quindi riferimento ai modelli PAM, *Piece-wise Exponential Additive Model*). Nonostante il modello PAM modelli il rischio di base utilizzando una funzione non lineare, il tasso di rischio stimato rimane sempre costante negli intervalli.

Un modo comune per specificare ignote funzioni di lisciamento $f(X_{\cdot k}, t_j)$ è di utilizzare le splines, rappresentate da una somma pesata di M funzioni di base (Ruppert, Wand and Carroll, 2003).

Al fine di stimare il modello GAMM, utilizzando la modellazione per covariate dipendenti dal tempo, i dati richiedono una specifica struttura: gli episodi per ciascun soggetto vengono definiti in base al cambiamento di stato delle covariate, in maniera simile al modello di Cox (Bender and Scheipl, 2018).

4.5.1 Stima e confronto

Poiché nel modello di Cox i coefficienti relativi agli indici non risultano significativi, si vuole osservare se lisciando tali covariate, si colga un effetto sul rischio. Si pongono innanzitutto a confronto i coefficienti fissi stimati con il modello di Cox e con il modello PAM. In Tabella 4.7 sono riportano le stime, assieme all'intervallo di confidenza al 95% ottenuto con $\hat{\beta} \pm 1.96_{z_{0.975}\hat{se}}$:

Modello	**Variabile**	**Coefficiente**	**IC: estremo inferiore**	**IC: estremo superiore**
PAM con h	Sesso: M	-0.127	-0.306	0.052
Cox-PH con h	Sesso: M	0.377	0.196	0.558
PAM con u	Sesso: M	-0.137	-0.317	0.042
Cox-PH con u	Sesso: M	0.363	0.181	0.545
PAM con h	Area: Geoscienze	-0.016	-0.235	0.204
Cox-PH con h	Area: Geoscienze	-0.446	-0.686	-0.207
PAM con u	Area: Geoscienze	-0.080	-0.292	0.132
Cox-PH con u	Area: Geoscienze	-0.493	-0.708	-0.277
PAM con h	Area: Statistica	0.333	0.117	0.548
Cox-PH con h	Area: Statistica	0.390	0.183	0.598
PAM con u	Area: Statistica	0.535	0.317	0.753
Cox-PH con u	Area: Statistica	0.415	0.204	0.625
PAM con h	Anno di inizio carriera	0.0270	0.0130	0.041
Cox-PH con h	Anno di inizio carriera	0.058	0.0435	0.073
PAM con u	Anno di inizio carriera	0.009	-0.005	0.024
Cox-PH con u	Anno di inizio carriera	0.056	0.04	0.071

Tabella 4.6: Stima e intervalli di confidenza al 95% per i coefficienti fissi

Si osserva innanzitutto che le stime dei coefficienti ottenuti dal modello di Cox, con u o con h sono molto vicine tra loro. Diversamente, il modello PAM stimato con u o con h fornisce stime dei coefficienti che differiscono tra loro, in particolare per le covariate relative all''Area di ricerca' e all''Anno di inizio carriera'. Il coefficiente relativo al 'Sesso' per il modello PAM, sia con u che con h, è inoltre negativo, al contrario di quanto osservato per il modello di Cox. Tuttavia nel modello PAM stimato sia con l'indice u o con l'indice h, gli intervalli di confidenza al 95% includono lo zero e valori positivi, seppure di poco. Il coefficiente relativo all''Area di ricerca' di Geoscienze è più vicino allo zero per il modello PAM rispetto al modello di Cox: l'intervallo di confidenza per il modello PAM è più stretto rispetto al modello di Cox e inoltre include lo zero, diversamente da quello per il modello di Cox.

In Figura 4.10 e 4.11 si riportano le stime dei coefficienti e i relativi intervalli di confidenza al 95%, ottenuti con $\hat{\beta} \pm 1.96_{z_{0.975}\hat{se}}$:

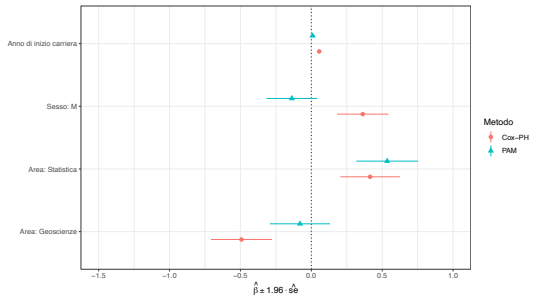

Figura 4.10: Stime e IC al 95% dei coefficienti fissi per i modelli Cox e PAM contenenti l'indice u

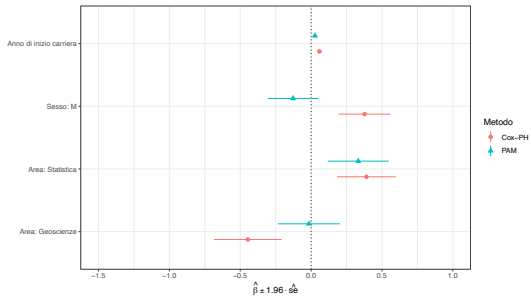

Figura 4.11: Stime e IC al 95% dei coefficienti fissi per i modelli Cox e PAM contenenti l'indice h

Si osserva quindi, in generale, che per entrambi i modelli PAM stimati, con l'indice u o con l'indice h, i coefficienti fissi hanno un effetto sul rischio inferiore rispetto ai coefficienti dei modelli di Cox stimati con u o con h.

Si considerino ora le due covariate dipendenti dal tempo: l'indice u e l'indice h. Nella stima dei modelli PAM, analogamente a quanto fatto per i modelli di Cox, si sono considerati u e h, separatamente. In Figura 4.12 è riportato l'effetto sul rischio del coefficiente relativo a u, in Figura 4.13 l'effetto di h sul rischio, per il modello PAM e per il modello di Cox:

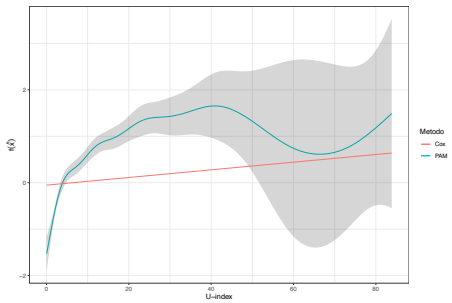

Figura 4.12: Effetto indice u sul rischio

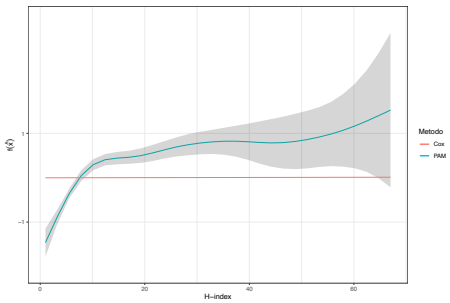

Figura 4.13: Effetto indice h sul rischio

L'andamento dell'indice u nel tempo è crescente, sia nel modello PAM che nel modello di Cox, anche se in maniera diversa. Nonostante il coefficiente relativo a u non risulti significativo nel modello di Cox, si osserva che l'effetto sul rischio, seppur lieve, è presente. In particolare l'effetto del coefficiente sul rischio cresce, in maniera lineare, all'aumentare dei valori dell'indice. Per valori dell'indice superiori a 50 l'effetto del coef-

ficiente stimato con il modello di Cox è compreso nelle bande di confidenza della stima ottenuta col modello PAM. L'effetto del coefficiente u sul rischio, stimato col modello PAM, cresce molto rapidamente per valori dell'indice compresi tra 0 e 10. Un valore dell'indice pari a 20, non influisce tuttavia sul rischio in maniera molto diversa da un valore pari a 40. Si osserva inoltre che per valori degli indici da 40 a 60, l'effetto del coefficiente sul rischio è decrescente e le bande di variabilità sono inoltre più ampie per valori superiori a 40.
Come precedentemente osservato in Tabella 4.5, il coefficiente relativo all'indice h, per il modello di Cox, non risulta significativo, considerando la sua evoluzione nel tempo. Notevolmente diverso è l'effetto dell'indice se modellato con una funzione di lisciamento. Il coefficiente relativo ad h, opportunamente lisciato attraverso le *splines*, mostra un effetto significativo nel tempo. In particolare l'effetto del coefficiente sul rischio cresce più rapidamente nell'intervallo di valori (0,10), più lentamente per valori superiori. Un valore dell'indice pari a 10 o pari a 25 fa sì che l'effetto di h sul rischio sia circa lo stesso.

Si riportano in Tabella 4.7 le stime di AIC e BIC per il modello stimato usando l'indice h e per il modello stimato usando l'indice u, con il metodo PAM e con il modello di Cox:

Criterio di confronto	**Modello**	**Modello con u-index**	**Modello con h-index**
AIC	PAM	5525	5642
	Cox	6953	6955
BIC	PAM	5549	5670
	Cox	6989	6989

Tabella 4.7: Risultati del confronto dei modelli

I criteri di selezione automatica sono notevolmente inferiori per il modello PAM, rispetto al modello di Cox. Sia AIC che BIC, per il modello PAM, mostrano nettamente che il modello con l'indice u è superiore rispetto al modello stimato con l'indice h. Per il modello di Cox, secondo l'AIC il modello migliore risulta essere quello contenente l'indice u, secondo il BIC quello contenente l'indice h, nonostante la differenza risulti minima. I valori dei criteri di selezione automatica osservati per il modello di Cox sono inoltre vicini a quelli ottenuti dal modello di Cox, stimato utilizzando i valori degli indici a diversi stadi di carriera (Capitolo 4, Tabella 4.3).

4.6 Foreste casuali di sopravvivenza

I modelli non parametrici non fanno assunzioni a priori sui dati. Essi sono diventati popolari nella costruzione di modelli di rischio predittivi nell'analisi di sopravvivenza soprattutto per il fatto che non richiedono assunzioni di proporzionalità, come nel modello di Cox.

L'approccio delle foreste casuali di sopravvivenza (*Random Survival Models*, RSF) è stato proposto da Ishwaran et al. (2008). L'implementazione di tale modello segue la teoria delle foreste casuali (Breiman, 2001) estendendola ai dati di sopravvivenza: la presenza della censura è la caratteristica che complica alcuni aspetti nella messa in pratica. In particolare la regola di 'taglio' per far crescere gli alberi deve tenere in considerazione la durata dell'evento, al fine di considerare se questa sia nota con certezza o censurata. La regola di suddivisione sulla quale si basa la costruzione della foresta è il test *log-rank*. Tale test, tradizionalmente usato per effettuare test a due campioni sui dati di sopravvivenza, può anche essere utilizzato nella suddivisione della sopravvivenza, per massimizzare le differenze tra i nodi.

Si consideri uno specifico nodo dell'albero che dev'essere suddiviso. Senza perdita di generalità si assuma che questo sia il nodo iniziale: il percorso dell'albero è definito dai dati $(T_1, X_1, \delta_1), ..., (T_n, X_n, \delta_n)$. Sia X una variabile specifica (ad esempio nominale) utilizzata per la suddivisione in nodi figli di sinistra e di destra, rispettivamente della forma $L = \{X_i \leq c\}$ e $R = \{X_i > c\}$. Siano $t_1 < t_2 < ... < t_m$ le durate distinte (in anni) in cui si verifica l'evento; $d_{j,L}$, $d_{j,R}$ e $Y_{j,L}, Y_{j,R}$ il numero di eventi accaduti e individui a rischio al tempo t_j nei nodi figli L, R.

Si considerino a rischio il numero di individui all'interno di un nodo figlio che non hanno ancora sperimentato l'evento o lo sperimentano al tempo t_j:

$$Y_{j,L} = \{T_i \geq t_j, X_i \leq c\}, \qquad Y_{j,R} = \{T_i \geq t_j, X_i > c\}. \tag{4.15}$$

Indicando con

$$Y_j = Y_{j,L} + Y_{j,R}, \qquad d_j = d_{j,L} + d_{j,R}, \tag{4.16}$$

il valore della statistica *log-rank* per la suddivisione è pari a:

$$L(X,c) = \frac{\sum_{j=1}^{m}(d_{j,L} - Y_{j,L}\frac{d_j}{Y_j})}{\sqrt{\sum_{j=1}^{m}\frac{Y_{j,L}}{Y_j}(1-\frac{Y_{j,L}}{Y_j})(\frac{Y_j-d_j}{Y_j-1})d_j}} \tag{4.17}$$

Il valore $|L(X,c)|$ è la misura di separazione del nodo. Più grande il valore, più grande la differenza di sopravvivenza tra L e R, e migliore la suddivisione. La miglior suddivisione è determinata individuando la covariata X e il valore di suddivisione c tale che $|L(X,c) \geq |L(X,c)|$ per ogni X e per ogni c.

Una volta che l'albero è stato fatto crescere, il predittore è definito all'interno di ciascun nodo terminale (ultimo nodo dell'albero). Indicando con h il nodo terminale, $t_{1,h} < t_{2,h} < ... < t_{m(h),h}$ le durate di carriera in cui si verifica l'evento in h, $d_{j,h}$ e $Y_{j,h}$ il numero di eventi e individui a rischio al tempo $t_{j,h}$, la funzione di sopravvivenza e la funzione di rischio sono definite nel seguente modo:

$$H_h(t) = \sum_{t_{j,h}\leq t}\frac{d_{j,h}}{Y_{j,h}}, \qquad S_h(t) = \prod_{t_{j,h}\leq t}(1-\frac{d_{j,h}}{Y_{j,h}}). \tag{4.18}$$

Le due quantità sono calcolate utilizzando gli stimatori bootstrap di Nelson-Aalen e Kaplan-Meier. Il predittore dell'albero di sopravvivenza è definito assegnando lo stesso rischio e la stessa stima della sopravvivenza a tutti i casi in h. Ciò si verifica perché lo scopo dell'albero di sopravvivenza è quello di partizionare i dati in gruppi omogenei di individui con comportamenti simili. Per la stima di $H(t|X)$ e $S(t|X)$ si utilizzano:

$$H(t|X) = H_h(t), \qquad S(t|X) = S_h(t), \qquad \text{se } X \in h \tag{4.19}$$

4.6.1 Stima del modello

Gli alberi che caratterizzano le foreste in questo caso sono costruiti utilizzando tecniche di ricampionamento: il predittore è caratterizzato dalla combinazione dei risultati di molti alberi. La strategia è la seguente:

- **Step 1.** Costruire B campioni bootstrap sui dati originali. Ogni campione bootstrap contiene circa 2/3 dei dati originali: la parte esclusa costituisce i cosiddetti dati fuori dal campione (*out-of-bag*, OOB);

- **Step 2.** Far crescere un albero basandosi sui dati di ciascun campione bootstrap $b = 1, ..., B$:

  1. per ciascun nodo selezionare un sottoinsieme di variabili predittive per la suddivisione;
  2. tra tutte le suddivisioni binarie definite dalle variabili predittive in (1) trovare la miglior suddivisione nei due sottoinsiemi che massimizzi le differenze di sopravvivenza tra nodi figli, utilizzando un criterio appropriato come il test *log-rank*;
  3. ripetere ricorsivamente (1) - (2) su ciascun nodo figlio fino a che non si raggiunge un criterio di arresto.

- **Step 3.** Calcolare la funzione di rischio cumulato e la funzione di sopravvivenza per ogni albero. La media di tutti gli alberi che sono stati fatti crescere fornisce la funzione di rischio cumulato totale.

- **Step 4.** Utilizzando i dati esclusi nello **Step 1.**, l'errore di previsione viene calcolato per la funzione di rischio cumulato totale e viene fornita una misura di importanza delle variabili presenti nel modello.

Come per il modello di Cox, utilizzando i tre insiemi di osservazioni contenenti i docenti e i loro indici a uno, cinque e dieci anni dalla prima pubblicazione, si stimano sei diversi modelli, contenenti ciascuno l'indice u o l'indice h. Il tasso di errore e l'importanza delle variabili dai modelli stimati è riportato nelle Figure 4.14, 4.15, 4.16, 4.17, 4.18, 4.19:

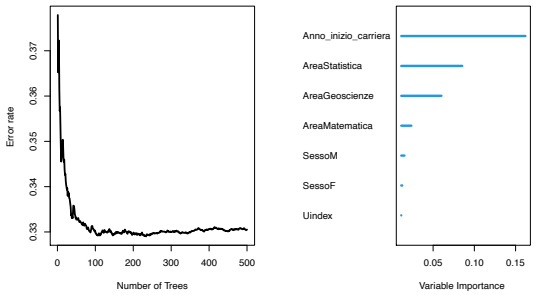

Figura 4.14: Foresta casuale stimata usando l'indice u a 1 anno da inizio carriera

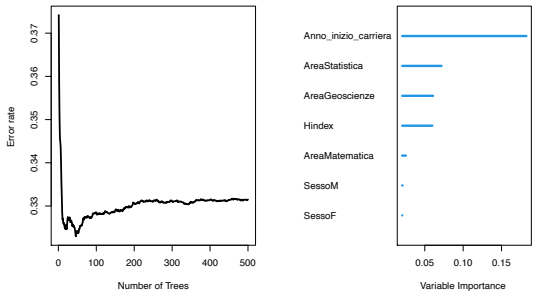

Figura 4.15: Foresta casuale stimata usando l'indice h a 1 anno da inizio carriera

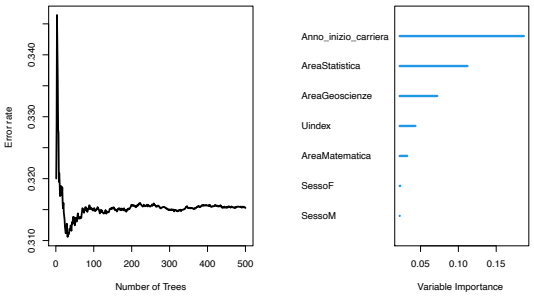

Figura 4.16: Foresta casuale stimata usando l'indice u a 5 anni da inizio carriera

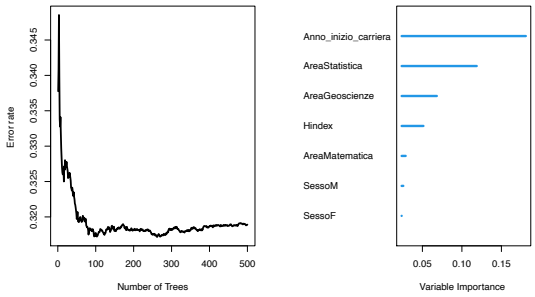

Figura 4.17: Foresta casuale stimata usando l'indice h a 5 anni da inizio carriera

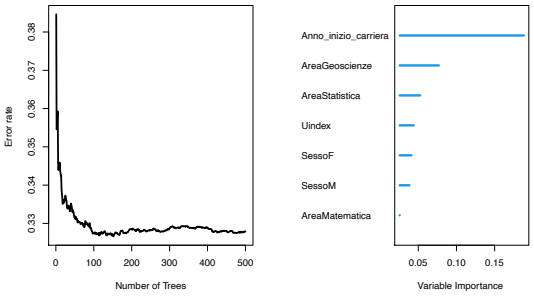

Figura 4.18: Foresta casuale stimata usando l'indice u a 10 anno da inizio carriera

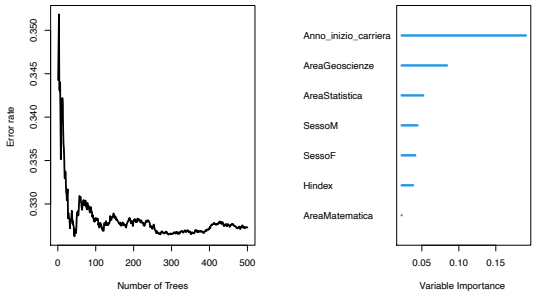

Figura 4.19: Foresta casuale stimata usando l'indice h a 10 anni da inizio carriera

Trattandosi di un metodo non parametrico, esso non fornisce una stima dei coefficienti ma soltanto un'indicazione relativa all'importanza delle variabili. Si nota che in tutti e tre gli insiemi di dati considerati, le variabili relative agli indici non hanno un forte impatto nel modello. L'*h-index* sembra essere più importante rispetto a u a un anno dalla prima pubblicazione; tuttavia l'importanza relativa decresce se considerato a 5 e 10 anni dall'inizio di carriera. Si osserva inoltre che stimando le foreste con un numero di alberi inferiore a 100, il tasso di errore si stabilizza. Inoltre quest'ultimo risulta essere inferiore per i modelli stimati utilizzando i valori degli indici a 5 anni da inizio carriera.

In Tabella 4.8 si riporta l'importanza relativa degli indici, nei modelli stimati utilizzando con h o u a uno, cinque, dieci anni di carriera:

Modello con indice	**Importanza relativa**
U a 1 anno	0.07
H a 1 anno	0.33
U a 5 anni	0.23
H a 5 anni	0.28
U a 10 anni	0.23
H a 10 anni	0.20

Tabella 4.8: Importanza relativa degli indici nelle foreste di sopravvivenza

Dai risultati ottenuti, le foreste casuali di sopravvivenza mostrano che il coefficiente relativo ad h ha un'importanza relativa maggiore nei primi anni di carriera. Dopo il quinto anno dalla prima pubblicazione, l'importanza relativa dell'indice u aumenta, quella dell'indice h diminuisce.

4.6.2 Previsione e confronto

Una misura di confronto in termini di accuratezza predittiva è data dal *Brier score* (Gerds, Cai and Schumacher, 2008). Tale misura, nata originariamente per valutare l'incertezza di un classificatore, è definita come la differenza quadratica media tra le probabilità previste e le osservazioni effettive:

$$BS(t, \hat{S}) = \mathrm{E}(S_i(t) - \hat{S}(t|X_i))^2 \tag{4.20}$$

con $i = 1, ..., N$ si indica la dimensione del campione, $S_i(t)$ rappresenta il vero stato del soggetto i al tempo t, $\hat{S}(t|X_i)$ la probabilità di sopravvivenza prevista al tempo t per il soggetto i con variabili predittive X_i.

Il *Brier score* ($BS(t)$) per i dati di sopravvivenza è definito in funzione del tempo:

$$\hat{BS}(t) = \frac{1}{N} \sum_{i=1}^{N} \left[0 - \hat{S}(t|X_i)\right]^2 \frac{\mathbb{1}(t_i \leq t, \delta_i = 1)}{\hat{G}(t_i|X_i)} + \left[1 - \hat{S}(t|X_i)\right]^2 \frac{\mathbb{1}(t_i > t)}{\hat{G}(t|X_i)} \tag{4.21}$$

dove $\hat{G}$ è la stima di Kaplan-Meier della funzione di sopravvivenza condizionata delle durate censurate. Il *Brier score* ha valori compresi tra 0 e 1. Una buona pre-

visione al tempo t è denotata da piccoli valori dello stesso. Utili valori di riferimento sono inoltre il 33%, che corrisponde a una previsione del rischio casuale (estraendo un numero da un uniforme $U[0,1)$, e il 25% che corrisponde a una previsione del 50% di rischio per ogni singolo soggetto.

Le curve dell'errore di previsione sono riportate nelle Figure 4.20, 4.21 e 4.22. Il modello rappresentato dalla curva nera è il modello di 'riferimento', privo di covariate. Rispetto a questo, si nota che entrambi i modelli contenenti gli indici u o h apportano una riduzione dell'errore di previsione. Le differenze maggiori tra il modello privo di covariate e i due modelli contenenti gli indici, si osservano utilizzando gli indici al primo anno di carriera, in Figura 4.20:

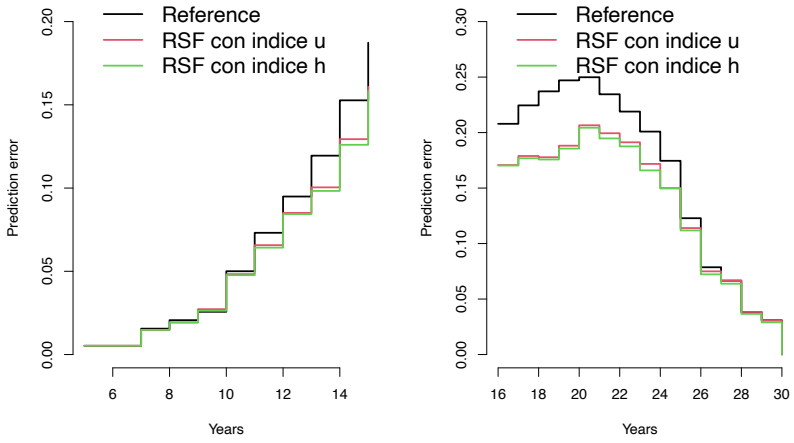

Figura 4.20: Errore di previsione al variare degli anni con indici a un anno da inizio carriera

Si osserva la maggiore capacità predittiva dell'indice h rispetto a u, in corrispondenza di durate di carriera > 14 anni. Tuttavia l'errore di previsione più piccolo si ha nelle durate di carriera più brevi, dove i modelli stimati con u o h forniscono simili errori.

Meno evidenti sono le differenze tra il modello privo di covariate e i modelli contenenti u o h, considerando gli indici a cinque e dieci anni dalla prima pubblicazione, come mostrato nelle Figure 4.21 e 4.22:

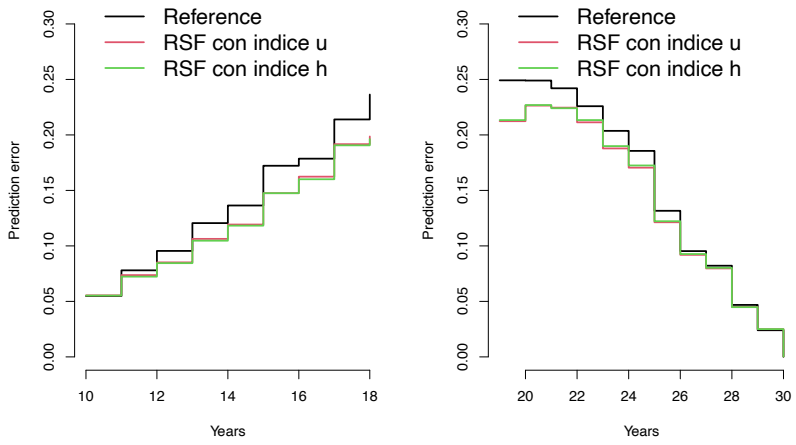

Figura 4.21: Errore di previsione al variare degli anni con indici a cinque anni da inizio carriera

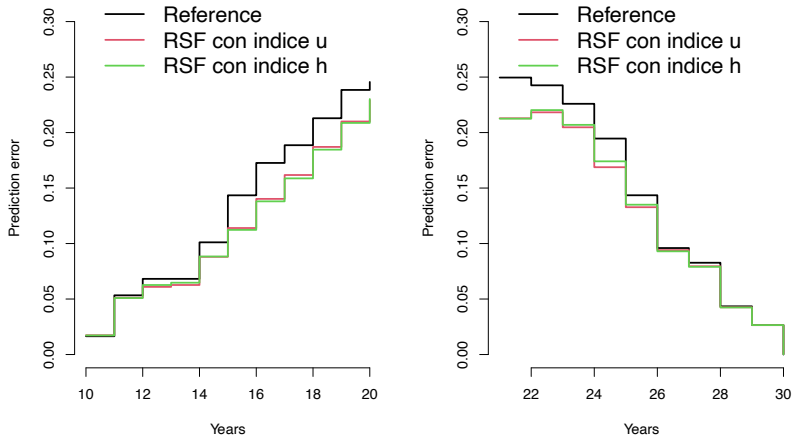

Figura 4.22: Errore di previsione al variare degli anni con indici a dieci anni da inizio carriera

Anche in questo caso l'errore di previsione è più piccolo quando le durate di carriera sono brevi. In corrispondenza del valore mediano di durata di carriera, 19 anni, l'errore è circa 0.25 per il modello privo di covariate, compreso tra 0.15 e 0.20 per il modello con u o con h, considerando gli indici a cinque o dieci anni dalla prima pubblicazione. Il modello contenente h fornisce un errore leggermente inferiore rispetto al modello con u,

in corrispondenza di durate di carriera brevi, viceversa per durate di carriera maggiori. Tale caratteristica mostra dunque migliori capacità in termini predittivi dell'indice h nel breve periodo, e di u nel lungo periodo.

L' *Integrated Brier Score* (IBS, Gerds and Schumacher (2006)) fornisce una misura complessiva delle prestazioni del modello per ogni istante di tempo $t_1 \leq t \leq t_{max}$ e può essere utilizzato per confrontare le capacità predittive dei modelli, a diversi istanti di tempo. L'IBS, dipendente dal tempo, nell'intervallo $[t_1; t_{max}]$ è definito come:

$$IBS = \int_{t_1}^{t_{max}} \mathrm{BS}(t) dw(t) \tag{4.22}$$

dove la funzione peso è $w(t) = \frac{t}{t_{max}}$. L'IBS può essere stimato sostituendo $\hat{BS}$(t) in Equazione 4.21 a BS(t).
Si consideri l'IBS per il modello di Cox e per le foreste di sopravvivenza. In Tabella 4.9 si riportano i risultati della metrica stimata utilizzando i modelli contenenti gli indici a diversi stadi di carriera. L'IBS è stato calcolato in diversi istanti di tempo, a 10, 20 e 30 anni di carriera:

Modello	**IBS (10)**	**IBS (20)**	**IBS (30)**
RSF con u-index a 1 anno	0.007	0.069	0.087
RSF con h-index a 1 anno	0.007	0.068	0.086
Cox con u-index a 1 anno	0.007	0.078	0.095
Cox con h-index a 1 anno	0.007	0.079	0.096
RSF con u-index a 5 anni	0.005	0.07	0.093
RSF con h-index a 5 anni	0.005	0.07	0.093
Cox con u-index a 5 anni	0.004	0.077	0.1
Cox con h-index a 5 anni	0.006	0.085	0.107
RSF con u-index a 10 anni	-	0.055	0.083
RSF con h-index a 10 anni	-	0.054	0.084
Cox con u-index a 10 anni	-	0.064	0.093
Cox con h-index a 10 anni	-	0.064	0.093

Tabella 4.9: *Brier score* per il modello di Cox e le foreste casuali di sopravvivenza

Innanzitutto si osserva che i valori dell'IBS sono in generale più bassi per le foreste casuali rispetto al modello di Cox. In generale si nota che più l'orizzonte temporale su cui si calcola il *Brier score* integrato è lontano dall'anno in cui si è rilevata la covariata,

più il BS aumenta. Considerando le covariate a 10 anni di carriera, sia per il modello di Cox che per le foreste di sopravvivenza, i valori dell'IBS sono inferiori rispetto ai valori ottenuti dai modelli contenenti gli indici a uno o cinque anni dalla prima pubblicazione. Utilizzando l'IBS come misura di confronto, i modelli stimati con l'approccio non parametrico sembrano dunque mostrare capacità predittive più elevate rispetto ai modelli semi-parametrici di Cox. Tuttavia le differenze non risultano così elevate, soprattutto stimando l'IBS in corrispondenza di durate di carriera ≤ 10 anni.

Capitolo 5

Analisi di sopravvivenza: classificazione di un gruppo di Ricercatori

La situazione più naturale che si incontra nello studio della sopravvivenza riguarda casi in cui un soggetto o gruppi di soggetti vengono seguiti nel tempo, fino alla realizzazione dell'evento o all'ultimo istante osservato, monitorando differenti aspetti di interesse. Le analisi condotte nel presente capitolo prendono in considerazione il campione di Ricercatori al 31/12/2005 selezionato nel Capitolo 2, esaminando gli anni di carriera impiegati per raggiungere il ruolo di Professore ordinario e valutando come gli indici, nel tempo, siano capaci di prevedere tale risultato.

5.1 Previsione delle carriere

Si consideri il campione selezionato a partire dai Ricercatori al 31/12/2005, costituito da 936 docenti e le loro relative $k = 14\,224$ pubblicazioni. Tra i Ricercatori in questione, al 31/12/2021 il 16% raggiunge il ruolo di Professore ordinario, il 45% ricopre il ruolo di Associato, il 39% di Ricercatore.

Poiché non tutti i docenti sperimentano l'evento di interesse entro la fine del periodo osservazionale è stata creata la variabile δ_i, per indicare se l'evento è stato osservato ($\delta_i = 1$) o censurato ($\delta_i = 0$) (a destra). Indicando con T_i la durata di carriera corrispondente al tempo in cui si verifica l'evento per il docente i, la curva di sopravvivenza

può essere rappresentata nel seguente modo:

$$S(t_i) = \Pr(T_i > t_i) \tag{5.1}$$

in cui $0 \leq S(t_i) \leq 1$ e $T_i \geq 0$.

Si consideri il modello di Cox, utilizzando la seguente equazione del rischio:

$$\lambda_i(t|X_i(t)) = \lambda_0(t)\exp(\beta_1^T x_{i1}(t) + ... + \beta_p^T x_{ip}(t)) \tag{5.2}$$

dove $\lambda_i(t|X_i(t))$ è il tasso di incidenza dell'evento al tempo t, λ_{0_t} rappresenta il rischio di base, β il coefficiente di regressione e X_i l'insieme di covariate.

5.2 Stima e confronto dei modelli

Dalle considerazioni effettuate al termine del Capitolo 3, volendo porre a confronto l'indice h e u, il campione selezionato a partire dai Ricercatori al 31/12/2005, costituito da 936 docenti e le loro relative $k = 14\,224$ pubblicazioni, è stato suddiviso. Si sono considerati tre diversi insiemi di osservazioni, nel seguente modo.
Per ciascun docente si sono considerate le informazioni contenute nelle variabili 'Anni di carriera', 'Sesso', 'Area di ricerca', 'Anno di inizio carriera', 'Evento', in corrispondenza dell'ultima pubblicazione disponibile per ciascuno. Si sono costruiti tre insiemi di osservazioni, utilizzando le covariate appena citate e estraendo i valori degli indici dal dataset di 14 224 pubblicazioni, in corrispondenza di tre durate di carriera diverse, per ciascuno:

- il primo insieme è caratterizzato da tutti i 936 docenti con valori degli indici a *un anno* da inizio carriera;
- il secondo insieme è caratterizzato da 925 docenti con valori degli indici a *cinque anni* da inizio carriera, eliminando i docenti che hanno raggiunto il ruolo di ordinario entro i 5 anni dalla prima pubblicazione;
- il terzo insieme è caratterizzato da 901 docenti con valori degli indici a *dieci anni* da inizio carriera, eliminando i docenti che hanno raggiunto il ruolo di ordinario

entro i 10 anni dalla prima pubblicazione;

I tre insiemi di osservazioni sono stati utilizzati nella fase di modellazione, dove l'indice h e l'indice u sono stati utilizzati separatamente, assieme alle altre covariate, nella stima dei modelli. In questo modo i modelli ottenuti sono stati confrontati sia in termini di capacità predittive fornite da ciascuno dei due indici, sia considerando le loro prestazioni a diversi stadi di carriera.

I modelli di Cox sono stati adattati sui tre insiemi di osservazioni in questione, utilizzando l'indice u o l'indice h, e confrontati seguendo la stessa metodologia esposta nel Capitolo 4, Sezione 4.3.2.
I modelli sono stati quindi così adattati:

- su una porzione di dati e confrontati attraverso tramite log-verosimiglianza predittiva ('Stima e verifica') e *Brier score*;
- su tutti i dati e confrontati tramite i criteri di selezione automatica ('Tutti i dati').

Le stime relative agli indici sono riportate in Tabella 5.1:

Modello con indice	**Stima e verifica**					**Tutti i dati**				
	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	z	$\Pr(>\|z\|)$	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	z	$\Pr(>\|z\|)$
U a 1 anno	0.778	2.177	0.35	2.22	0.026 *	0.539	1.714	0.324	1.661	0.097
H a 1 anno	0.341	1.406	0.104	3.289	0.001 **	0.242	1.273	0.096	2.524	0.012 *
U a 5 anni	0.628	1.874	0.092	6.819	9.14e-12 ***	0.666	1.946	0.072	9.287	$<$ 2e-16 ***
H a 5 anni	0.188	1.207	0.032	5.814	6.10e-09 ***	0.217	1.243	0.026	8.421	$<$ 2e-16 ***
U a 10 anni	0.283	1.327	0.028	10.059	$<$ 2e-16 ***	0.285	1.33	0.024	11.668	$<$ 2e-16 ***
H a 10 anni	0.114	1.12	0.012	9.452	$<$ 2e-16 ***	0.121	1.129	0.01	11.612	$<$ 2e-16 ***

Tabella 5.1: Stime dei coefficienti relativi agli indici utilizzando il modello di Cox

Dai risultati ottenuti si può osservare che le covariate relative agli indici hanno un effetto significativo nello spiegare il fenomeno di interesse. Il modello conferma che avere un valore più elevato dell'indice, al netto delle altre variabili, aumenta il rischio di sperimentare l'evento di interesse. Per entrambi gli indici, tale effetto sul rischio risulta maggiore nei primi anni di carriera, rispetto ai 5 e 10 anni dalla prima pubblicazione. Le stime ottenute dai modelli stimati utilizzando tutti i dati o il solo insieme di stima sono molto simili, così come gli errori standard. In particolare si osserva che considerando gli indici a uno, cinque, dieci anni di carriera, l'errore standard delle stime diminuisce.

In Tabella 5 si riportano i risultati di:

- log-verosimiglianza predittiva e *Brier score*, calcolati sulla porzione di dati esclusa nelle stime dei modelli;
- criteri di selezione automatica calcolati su tutti i dati.

Metodo di confronto	**Modello con indice**					
	U a 1 anno	**H a 1 anno**	**U a 5 anni**	**H a 5 anni**	**U a 10 anni**	**H a 10 anni**
PLL	-165.68	-165.88	-164.36	-167.13	-149.8	-150.2
AIC	1820	1817	1753	1765	1710	1712
BIC	1844	1841	1778	1790	1734	1736
IBS(0, 40)	0.043	0.042	0.034	0.035	0.027	0.028

Tabella 5.2: Risultati del confronto tra modelli

Si osserva che i valori della log-verosimiglianza predittiva forniti dal modello con l'indice u e con l'indice h non differiscono molto. Da quanto emerso risulta inoltre che considerando gli indici a 5 anni dalla prima pubblicazione le prestazioni in termini predittive fornite dai modelli sono migliori, rispetto agli altri due stadi degli indici considerati. I criteri di selezione automatica sottolineano maggiormente la differenza tra modelli stimati con u e con h: in corrispondenza del modello contenente gli indici a 5 e 10 anni da inizio carriera i valori minimi per AIC e BIC si hanno per il modello contenente l'indice u. Considerando le covariate a 1 anno dalla prima pubblicazione, il modello contenete l'indice h risulta essere lievemente migliore in termini predittivi.

In Figura 5.1 e in Figura 5.2 si riportano le curve relative al rischio cumulato per i due indici, selezionati a 5 anni da inizio carriera. Al fine della rappresentazione, si sono suddivisi i valori degli indici in 3 classi:

- u: classi [0, 1.5], (1.5,4], (4, 7];
- h: classi [1,3], (3,9], (9, 16].

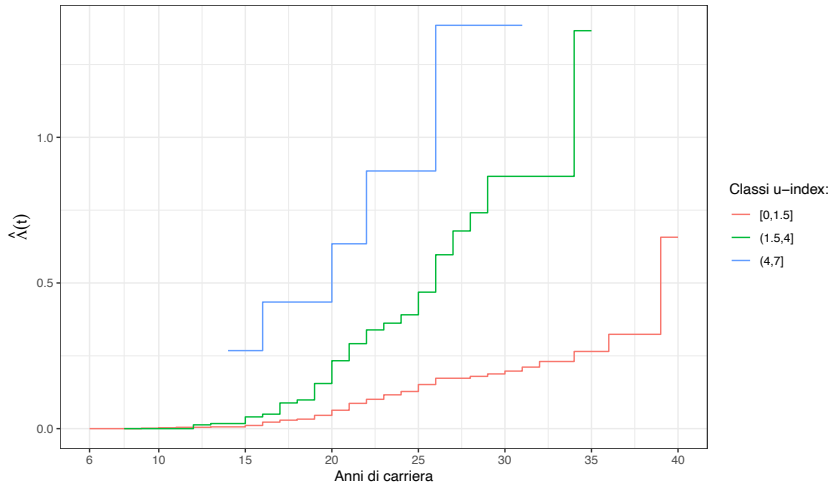

Figura 5.1: Rischio cumulato utilizzando i valori dell'indice u a cinque anni da inizio carriera

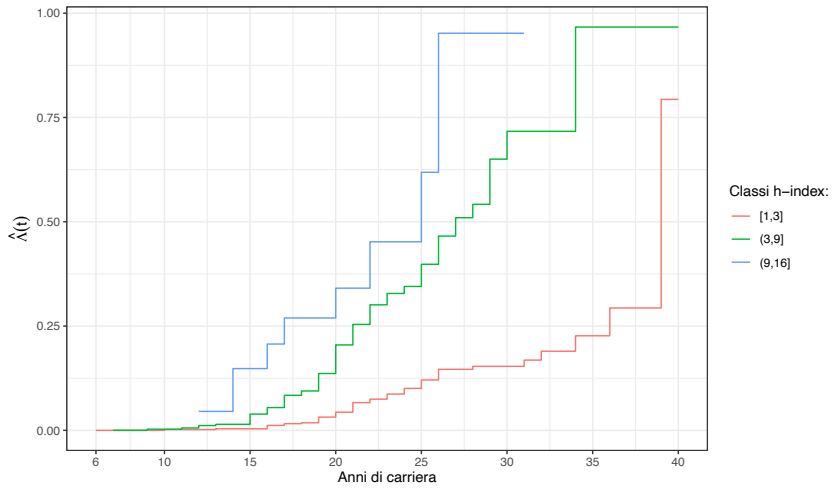

Figura 5.2: Rischio cumulato utilizzando i valori dell'indice h a cinque anni da inizio carriera

Quanto emerso dalle curve del rischio cumulato conferma le analisi effettuate in partenza: i Ricercatori con valori degli indici più elevati hanno un rischio maggiore di sperimentare l'evento rispetto a chi ha valori degli indici bassi. La differenza in termini di rischio di sperimentare l'evento tra chi ha valori degli indici molto elevati e nelle classi intermedie non risulta tuttavia così elevata.

5.3 Confronto predittivo: curva ROC

Quando una variabile risposta Y_i è binaria, l'accuratezza di una previsione o di una regola di classificazione è tipicamente riassunta attraverso il tasso di corretta classificazione definito come sensibilità, $P(p_i > c|Y_i = 1)$, e specificità, $P(p_i \leq c|Y_i = 0)$, dove p_i è una previsione, e c una soglia per classificare la previsione come positiva ($p_i > c$) o negativa ($p_i \leq c$). Quando un valore di c non è specificato a priori, sensibilità e specificità possono essere caratterizzate utilizzando la curva ROC che raffigura il 'tasso di veri positivi' (sensibilità) e il 'tasso di falsi positivi' (1-specificità) al variare di $c \in (-\infty, +\infty)$.

La curva ROC fornisce informazioni complete sull'insieme di tutte le possibili combinazioni dei tassi di veri positivi e falsi positivi, ma risulta anche utile come rappresentazione grafica della separazione tra distribuzioni di 'casi' e 'controlli'. Se tra tali due gruppi non vi è alcuna sovrapposizione nella distribuzione, la curva ROC assume valore 1 (tasso di veri positivi perfetto) per un qualsiasi tasso di falsi positivi maggiore di zero: la discriminazione tra casi e controlli risulta perfetta. Al contrario, se le due distribuzioni sono identiche la curva ROC starà sulla linea dei 45° indicando che la discriminazione è pessima. L'area sotto la curva (AUC, *Area Under the Curve*), rappresenta una misura di concordanza tra l'indicatore in esame e lo stato dei soggetti in esame (Hanley and McNeil, 1982).

Generalizzando i concetti di sensibilità e specificità per le applicazioni ai dati di sopravvivenza, Heagerty and Zheng (2005) hanno proposto diverse definizioni per stimare la sensibilità e specificità nel caso di dati censurati. Tra queste, la sensibilità incidentale e specificità dinamica permettono di considerare un soggetto come 'controllo' per un primo periodo, e come 'caso' in un qualche istante di tempo successivo.

Sia T_i la durata dell'evento, si definisce sensibilità incidentale la probabilità che un docente abbia un rischio p_i maggiore di una certa soglia c tra tutti i docenti che hanno sperimentato l'evento al tempo t e specificità dinamica la probabilità che un docente abbia un rischio p_i minore o uguale a c, tra tutti i docenti a rischio di sperimentare

l'evento al tempo t. La sensibilità, specificità e l'AUC(t) sono definiti come:

$$\begin{aligned} \textit{sensibilità}(c,\ t) &= P(p_i > c | T_i = t) \\ \textit{specificità}(c,\ t) &= P(p_i \leq c | T_i > t) \\ \mathrm{A}UC(t) &= P(p_i > p_j | T_i = t, T_j > t), \quad i \neq j \end{aligned} \tag{5.3}$$

dove p_i rappresenta il rischio e c una soglia. Si osservi che p_i potrebbe anche essere sostituito con un valore di una covariata.
La sensibilità incidentale e specificità dinamica sono definite dicotomizzando l'insieme di rischio al tempo t, tra i soggetti che hanno (casi) e non hanno (controlli) sperimentato l'evento. Tale definizione di specificità e sensibilità risulta appropriata nel presente contesto poiché la durata in corrispondenza della quale si verifica l'evento è nota e si vogliono discriminare i docenti che sperimentano l'evento da quelli che non lo subiscono entro la fine del periodo osservazionale.

Si considerino i modelli di Cox stimati nel paragrafo precedente: si vuole valutare se l'accuratezza cambi nel tempo e in che modo utilizzando i due indici. Con i predittori lineari stimati dal modello di Cox, si costruiscono le curve ROC incidentali/dinamiche (I/D). In tabella 5.3 si riportano i valori dell'AUC ottenuti, utilizzando diversi orizzonti di previsione temporale:

Indice utilizzato	**AUC a 5 anni**	**AUC a 10 anni**	**AUC a 15 anni**
U a 1 anno di carriera	0.962	0.925	0.882
H a 1 anno di carriera	0.968	0.927	0.885
U a 5 anni di carriera	0.958	0.928	0.891
H a 5 anni di carriera	0.956	0.927	0.889
U a 10 anni di carriera	-	0.931	0.898
H a 10 anni di carriera	-	0.926	0.896

Tabella 5.3: AUC stimato a diversi orizzonti predittivi considerando gli indici a diversi stadi di carriera

Si osserva che più la covariata è lontana dall'orizzonte di previsione, più l'AUC diminuisce, sia per l'indice h che per l'indice u. Considerando orizzonti predittivi differenti, l'effetto di h risulta migliore considerando l'indice a un anno da inizio carriera. Con valori degli indici a cinque, dieci anni di carriera tuttavia, u fornisce

valori più elevati dell'AUC.

Le curve ROC sono riportate nelle Figure 5.3, 5.4, 5.5, 5.6, 5.7, 5.8:

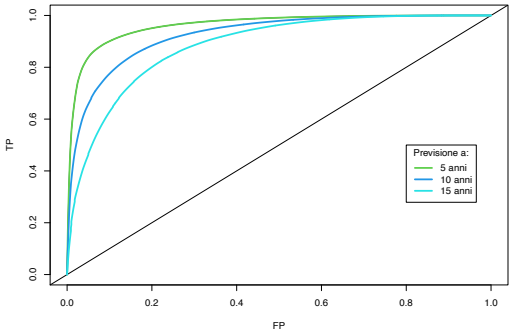

Figura 5.3: Curva ROC con *u-index* a 1 anno di carriera

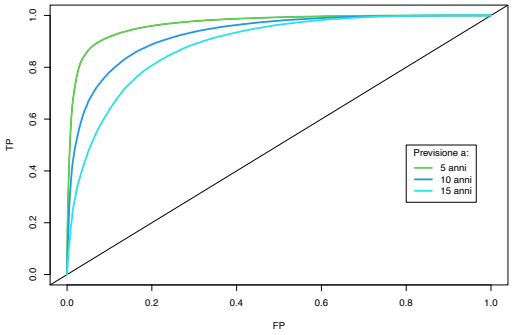

Figura 5.4: Curva ROC con *h-index* a 1 anno di carriera

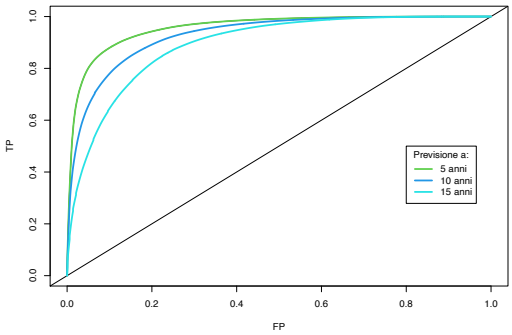

Figura 5.5: Curva ROC con *u-index* a 5 anni di carriera

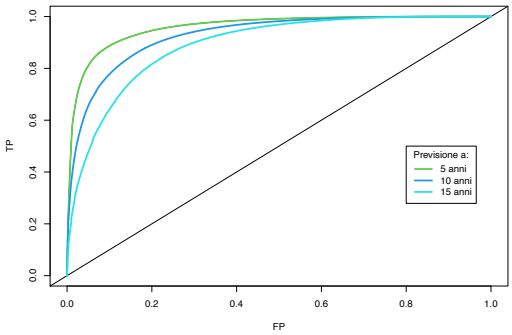

Figura 5.6: Curva ROC con *h-index* a 5 anni di carriera

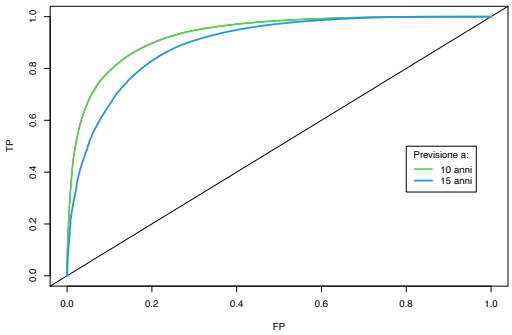

Figura 5.7: Curva ROC con *u-index* a 10 anni di carriera

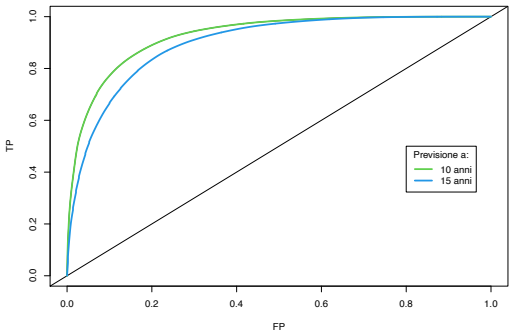

Figura 5.8: Curva ROC con *h-index* a 10 anni di carriera

Fissando un tasso predefinito di falsi positivi, in tutti i casi e per entrambi gli indici, si osserva che la sensibilità, ossia il tasso di veri positivi, è decrescente all'aumentare dell'orizzonte predittivo considerato. Le differenze tra i due indici in termini predittivi risultano minime.

Conclusione

Con questa tesi ci si è posti l'obiettivo di utilizzare gli indicatori bibliometrici come strumento predittivo della durata delle carriere dei docenti, fino al raggiungimento del ruolo di Professore ordinario, e confrontarne l'utilità come strumenti di valutazione.
Ponendo a confronto nelle analisi, l'indice Universale (*u-index*) e l'indice di Hirsch (*h-index*), si è osservato che entrambi possiedono buone capacità nel prevedere la risposta, in particolare se considerati nei primi dieci anni di carriera. Dalla stima del modello di Cox, per entrambi i campioni in esame, è emerso che l'effetto dell'indice u sul rischio è più forte rispetto a h, in tutti gli stadi di carriera considerati, a uno, cinque, dieci anni dalla prima pubblicazione. Considerando le due metriche di accuratezza utilizzate, dai valori ottenuti della log-verosimiglianza predittiva, i due indici sembrano avere capacità predittive simili. Tuttavia i criteri di selezione automatica, sottolineano che l'indice u è più predittivo di h, in particolare utilizzando i valori a cinque e dieci anni dalla prima pubblicazione. Utilizzando gli indici come covariate dipendenti dal tempo, si è osservato che la modellazione GAMM risulta più appropriata rispetto a quella di Cox, cogliendo l'effetto degli indici nel tempo. Questo aspetto è confermato dai criteri di selezione automatica, che mostrano inoltre una migliore capacità predittiva dell'indice u rispetto a h. L'approccio non parametrico delle foreste casuali di sopravvivenza sottolinea nuovamente che il coefficiente relativo all'indice h risulta più importante rispetto a u, al netto delle altre covariate inserite nel modello, tra il primo e il quinto anno di carriera. Tuttavia tale approccio non mostra differenze significative in termini predittivi utilizzando l'indice u o l'indice h. Anche valutando l'accuratezza delle previsioni nel tempo, mediante strumenti di classificazione dinamica, si è osservato che il modello contenente l'indice u fornisce valori di AUC più elevati rispetto a quello contenente h, considerando le covariate a cinque e dieci anni dalla prima pubblicazione.
Per entrambi i campioni utilizzati, di Professori ordinari e Ricercatori, utilizzando gli

indici bibliometrici a diversi stadi di carriera si è inoltre osservato che i valori degli stessi, condizionatamente ai periodi in esame, hanno un impatto differente sul rischio di sperimentare l'evento. Si è stimata l'incidenza dell'evento di interesse utilizzando la curva di sopravvivenza, ottenuta con lo stimatore di Kaplan-Meier, per i valori degli indici a uno, cinque e dieci anni di carriera. Nonostante i due indici siano costruiti utilizzando variabili diverse e di conseguenza presentino scale di misura diverse, per entrambi si è osservato che docenti con valori molto elevati degli indici non hanno un rischio molto più alto di sperimentare l'evento di interesse, rispetto a docenti con valori degli indici intermedi. Tale effetto si nota maggiormente a partire dai valori a cinque anni dalla prima pubblicazione, dove gli indici cominciano a differenziarsi maggiormente.
Si può quindi affermare che entrambi gli indici sono buoni predittori delle carriere accademiche. L'indice h ha buone capacità nel prevedere le durate di carriera se considerato negli anni accademici iniziali. Tuttavia dal momento in cui non si registra un aumento del suo valore e l'indice rimane costante, le capacità predittive decrescono rispetto a quelle fornite dall'indice u, che si mantiene buon predittore per tutto l'arco di carriera. Inoltre quest'ultimo, tenendo conto di informazioni aggiuntive rispetto ad h, permette di monitorare con maggiore precisione l'impatto bibliometrico di un autore, risultando quindi più completo se utilizzato come strumento di valutazione. L'indice h rimane comunque più semplice da calcolare e facile da interpretare, necessitando di sole due informazioni: le pubblicazioni e il relativo numero di citazioni.

Appendice

Nella Tabella 4 vengono riportati i risultati dei modelli di Cox stimati sul campione selezionato a partire dai Professori ordinari al 31/12/2021. Si riportano le stime ottenute utilizzando il metodo della verosimiglianza parziale ('Cox') e verosimiglianza parziale penalizzata ('Cox p.'), per i modelli contenenti l'indice u o h a uno, cinque, dieci anni dalla prima pubblicazione.

Nella Tabella 5 vengono riportati i risultati dei modelli di Cox stimati sul campione selezionato a partire dai Ricercatori al 31/12/2005, contenenti l'indice u o h a uno, cinque, dieci anni dalla prima pubblicazione.

In Tabella 4 si riportano le stime delle covariate contenute nei modelli di Cox stimati nel Capitolo 4, Sezione 4.3:

Modello con	**Metodologia**	**Coefficiente**	**Stima e Verifica**					**Tutti i dati**				
			$\hat{\beta}$	$\exp(\hat{\beta})$	$se(\hat{\beta})$	z	$Pr(>\|z\|)$	$\hat{\beta}$	$\exp(\hat{\beta})$	$se(\hat{\beta})$	z	$Pr(>\|z\|)$
Indice u a 1 anno	Cox	Sesso: M	0.397	1.488	0.105	3.8	0.0002 ***	0.382	1.466	0.09	4.242	2.21e-05 ***
	Cox p.	Sesso: M	0.355	-	-	-	-	0.343	-	-	-	-
	Cox	Area: Geoscienze	-0.424	0.654	0.114	-3.708	0.0002 ***	-0.434	0.648	0.101	-4.320	1.56e-05 ***
	Cox p.	Area: Geoscienze	-0.39	-	-	-	-	-0.402	-	-	-	-
	Cox	Area: Statistica	0.33	1.359	0.123	2.682	0.007 **	0.419	1.521	0.105	4.007	6.15e-05 ***
	Cox p.	Area: Statistica	0.293	-	-	-	-	0.386	-	-	-	-
	Cox	Anno di inizio carriera	0.06	1.062	0.009	6.765	1.33e-11 ***	0.311	1.365	0.166	1.876	0.061
	Cox p.	Anno di inizio carriera	0.055	-	-	-	-	0.053	-	-	-	-
Indice h a 1 anno	Cox	Sesso: M	0.386	1.471	0.104	3.698	0.0002 ***	0.379	1.461	0.09	4.204	2.62e-05 ***
	Cox p.	Sesso: M	0.322	-	-	-	-	0.34	-	-	-	-
	Cox	Area: Geoscienze	-0.435	0.647	0.115	-3.798	0.0002 ***	-0.448	0.639	0.105	-4.460	8.18e-06 ***
	Cox p.	Area: Geoscienze	-0.381	-	-	-	-	-0.412	-	-	-	-
	Cox	Area: Statistica	0.3	1.35	0.122	2.455	0.014 *	0.391	1.478	0.103	3.779	0.0002 ***
	Cox p.	Area: Statistica	0.256	-	-	-	-	0.361	-	-	-	-
	Cox	Anno di inizio carriera	0.06	1.062	0.01	6.794	1.09e-11 ***	0.058	1.06	0.007	7.873	3.47e-156 ***
	Cox p.	Anno di inizio carriera	0.054	-	-	-	-	0.053	-	-	-	-
Indice u a 5 anni	Cox	Sesso: M	0.431	1.539	0.105	4.098	4.16e-05 ***	0.369	1.447	0.09	4.091	4.30e-05 ***
	Cox p.	Sesso: M	0.385	-	-	-	-	0.332	-	-	-	-
	Cox	Area: Geoscienze	-0.437	0.646	0.114	-3.842	0.0001 ***	-0.445	0.641	0.1	-4.437	9.11e-06 ***
	Cox p.	Area: Geoscienze	-0.41	-	-	-	-	-0.411	-	-	-	-
	Cox	Area: Statistica	0.656	1.927	0.124	5.275	1.33e-07 **	0.424	1.528	0.107	3.981	6.86e-05 ***
	Cox p.	Area: Statistica	0.597	-	-	-	-	0.389	-	-	-	-
	Cox	Anno di inizio carriera	0.054	1.055	0.009	6.306	2.87e-10 ***	0.057	1.058	0.008	7.495	6.61e-14
	Cox p.	Anno di inizio carriera	0.049	-	-	-	-	0.052	-	-	-	-
Indice h a 5 anni	Cox	Sesso: M	0.438	1.55	0.105	4.166	3.10e-05 ***	0.376	1.456	0.09	4.165	3.12e-05 ***
	Cox p.	Sesso: M	0.392	-	-	-	-	0.337	-	-	-	-
	Cox	Area: Geoscienze	-0.439	0.645	0.114	-3.855	0.0001 ***	-0.445	0.641	0.1	-4.436	9.16e-06 ***
	Cox p.	Area: Geoscienze	-0.411	-	-	-	-	-0.411	-	-	-	-
	Cox	Area: Statistica	0.624	1.867	0.123	5.07	3.97e-07 ***	0.4	1.486	0.105	3.771	0.0002 ***
	Cox p.	Area: Statistica	0.568	-	-	-	-	0.363	-	-	-	-
	Cox	Anno di inizio carriera	0.055	1.057	0.009	6.434	1.24e-10 ***	0.06	1.06	0.008	7.671	1.71e-14 ***
	Cox p.	Anno di inizio carriera	0.05	-	-	-	-	0.053	-	-	-	-
Indice u a 10 anni	Cox	Sesso: M	0.455	1.576	0.108	4.199	2.68e-05 ***	0.36	1.433	0.093	3.859	0.0001 ***
	Cox p.	Sesso: M	0.409	-	-	-	-	0.323	-	-	-	-
	Cox	Area: Geoscienze	-0.482	0.618	0.116	-4.152	3.29e-05***	-0.476	0.621	0.102	-4.664	3.10e-06 ***
	Cox p.	Area: Geoscienze	-0.446	-	-	-	-	-0.438	-	-	-	-
	Cox	Area: Statistica	0.393	1.481	0.131	2.989	0.003 **	0.312	1.366	0.115	2.715	0.007 ***
	Cox p.	Area: Statistica	0.345	-	-	-	-	0.275	-	-	-	-
	Cox	Anno di inizio carriera	0.042	1.043	0.009	4.785	1.71e-06 ***	0.05	1.0514	0.008	6.423	1.33e-10
	Cox p.	Anno di inizio carriera	0.039	-	-	-	-	0.046	-	-	-	-
Indice h a 10 anni	Cox	Sesso: M	0.463	1.589	0.108	4.276	1.90e-05 ***	0.369	1.446	0.093	3.961	7.47e-05 ***
	Cox p.	Sesso: M	0.417	-	-	-	-	0.332	-	-	-	-
	Cox	Area: Geoscienze	-0.479	0.62	0.117	-4.081	4.49e-05 ***	-0.461	0.631	0.103	-4.487	7.23e-06 ***
	Cox p.	Area: Geoscienze	-0.443	-	-	-	-	-0.425	-	-	-	-
	Cox	Area: Statistica	0.351	1.42	0.13	2.7	0.007 **	0.266	1.304	0.114	2.341	0.02 *
	Cox p.	Area: Statistica	0.306	-	-	-	-	0.233	-	-	-	-
	Cox	Anno di inizio carriera	0.045	1.046	0.009	5.093	3.52e-07 ***	0.053	1.055	0.008	6.781	1.19e-11 ***
	Cox p.	Anno di inizio carriera	0.041	-	-	-	-	0.048	-	-	-	-

Tabella 4: Stime dei coefficienti del modello di Cox per il campione di Professori ordinari al 31/12/2021

In Tabella 5 si riportano le stime dei coefficienti contenuti nei modelli di Cox del Capitolo 5, Sezione 5.1:

Modello con	**Coefficiente**	**Stima e Verifica**					**Tutti i dati**				
		$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	z	$\Pr(>\|z\|)$	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	z	$\Pr(>\|z\|)$
Indice u a un anno	Sesso: M	0.646	1.908	0.202	3.196	0.001 **	0.537	1.711	0.169	3.173	0.002 **
	Area: Matematica	-0.064	0.938	0.228	-0.281	0.779	-0.168	0.845	0.195	-0.863	0.388
	Area: Statistica	0.963	2.619	0.29	3.325	0.001 ***	0.763	2.145	0.243	3.136	0.002 **
	Anno di inizio carriera	0.287	1.332	0.041	7.081	1.43e-12 ***	0.31	1.363	0.035	8.820	< 2e-16 ***
Indice h a un anno	Sesso: M	0.672	1.957	0.203	3.307	0.001 ***	0.582	1.79	0.172	3.387	0.001 ***
	Area: Matematica	-0.06	1.062	0.232	0.259	0.796	-0.109	0.897	0.196	-0.555	0.579
	Area: Statistica	1.051	2.861	0.294	3.574	0.0004 ***	0.785	2.193	0.244	3.225	0.001 **
	Anno di inizio carriera	0.285	1.329	0.04	7.108	1.18e-12 ***	0.306	1.358	0.035	8.786	< 2e-16 ***
Indice u a 5 anni	Sesso: M	0.492	1.635	0.203	2.426	0.015 *	0.51	1.665	0.17	3.008	0.003 **
	Area: Matematica	0.08	1.083	0.232	0.344	0.731	-0.008	0.992	0.198	-0.041	0.967
	Area: Statistica	1.357	3.884	0.293	4.632	3.62e-06 ***	1.231	3.424	0.251	4.906	9.30e-07 ***
	Anno di inizio carriera	0.265	1.303	0.042	6.254	4.00e-10 ***	0.291	1.338	0.037	7.946	1.92e-15 ***
Indice h a 5 anni	Sesso: M	0.526	1.693	0.202	2.605	0.009 **	0.557	1.745	0.17	3.278	0.001 **
	Area: Matematica	0.111	1.118	0.233	0.478	0.633	0.007	1.007	0.198	0.037	0.971
	Area: Statistica	1.204	3.333	0.282	4.266	1.99e-05 ***	1.087	2.967	0.241	4.508	6.55e-06 ***
	Anno di inizio carriera	0.266	1.305	0.041	6.437	1.22e-10 ***	0.291	1.338	0.036	8.147	3.73e-16 ***
Indice u a 10 anni	Sesso: M	0.373	1.453	0.206	1.812	0.07	0.326	1.386	0.172	1.897	0.058
	Area: Matematica	0.488	1.63	0.233	2.099	0.036	0.158	1.172	0.201	0.79	0.43
	Area: Statistica	1.991	7.325	0.261	7.637	2.23e-14 ***	1.175	3.24	0.24	4.9	9.44e-07 ***
	Anno di inizio carriera	0.265	1.303	0.042	6.254	4.00e-10 ***	0.302	1.352	0.037	8.104	5.34e-16 ***
Indice h a 10 anni	Sesso: M	0.409	1.506	0.206	1.985	0.047 *	0.334	1.397	0.172	1.943	0.052
	Area: Matematica	0.831	2.296	0.247	3.371	0.0008 ***	0.53	1.7	0.208	2.543	0.011 *
	Area: Statistica	2.213	9.14	0.268	8.248	< 2e-16 ***	1.4	4.056	0.242	5.784	7.31e-09 ***
	Anno di inizio carriera	0.266	1.305	0.041	6.437	1.22e-10 ***	0.306	1.358	0.037	8.325	< 2e-16 ***

Tabella 5: Stime dei coefficienti del modello di Cox per il campione di Ricercatori al 31/12/2005

Bibliografia

Ajiferuke, I., Lu, K. and Wolfram, D. (2010), 'A comparison of citer and citation-based measure outcomes for multiple disciplines', *Journal of the American Society for Information Science and technology* .

Altman, D. G. and Bland, J. M. (1998), 'Time to event (survival) data', *Bmj* .

Asaolu, O. S., Jaiyeola, T. G., Usikalu, M. R., Gayawan, E., Atolani, O. and Adeyemi, O. S. (2022), 'U-index: A new universal metric as unique indicator of researcher's contributions to academic knowledge', *Scientific African* .

Bender, A., Groll, A. and Scheipl, F. (2018), 'A generalized additive model approach to time-to-event analysis', *Statistical Modelling* .

Bender, A. and Scheipl, F. (2018), 'Pammtools: Piece-wise exponential additive mixed modeling tools', *arXiv preprint arXiv:1806.01042* .

Bland, J. M. and Altman, D. G. (1998), 'Survival probabilities (the kaplan-meier method)', *Bmj* .

Bodapati, A. and Gupta, S. (2004), 'A direct approach to predicting discretized response in target marketing', *Journal of Marketing Research* .

Breiman, L. (2001), 'Random forests', *Machine learning* .

Burrell, Q. L. (2007), 'Hirsch's h-index: A stochastic model', *Journal of Informetrics* .

Cappelletti-Montano, B., Columbu, S., Montaldo, S. and Musio, M. (2021), 'New perspectives in bibliometric indicators: Moving from citations to citing authors', *Journal of Informetrics* .

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society. Series B (Methodological)* .

Dieks, D. and Chang, H. (1976), 'Differences in impact of scientific publications: Some indices derived from a citation analysis', *Social Studies of Science* .

Friedman, M. (1982), 'Piecewise exponential models for survival data with covariates', *The Annals of Statistics* .

Gerds, T. A., Cai, T. and Schumacher, M. (2008), 'The performance of risk prediction models', *Biometrical Journal: Journal of Mathematical Methods in Biosciences* .

Gerds, T. A. and Schumacher, M. (2006), 'Consistent estimation of the expected brier score in general survival models with right-censored event times', *Biometrical Journal* .

Gomez, G., Julià, O., Utzet, F. and Moeschberger, M. L. (1992), 'Survival analysis for left censored data', *Survival analysis: State of the art* .

Hanley, J. A. and McNeil, B. J. (1982), 'The meaning and use of the area under a receiver operating characteristic (roc) curve', *Radiology* .

Heagerty, P. J. and Zheng, Y. (2005), 'Survival model predictive accuracy and roc curves', *Biometrics* .

Hirsch, J. E. (2005), 'An index to quantify an individual's scientific research output', *Proceedings of the National Academy of Sciences of the United States of America* .

Ishwaran, H., Kogalur, U. B., Blackstone, E. H. and Lauer, M. S. (2008), 'Random survival forests'.

Kullback, S. and Leibler, R. A. (1951), 'On information and sufficiency', *The Annals of Mathematical Statistics* .

Lin, X. and Zhang, D. (1999), 'Inference in generalized additive mixed models by using smoothing splines', *Journal of the Royal Statistical Society Series B: Statistical Methodology* .

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003), *Semiparametric regression*.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011), 'Regularization paths for cox's proportional hazards model via coordinate descent', *Journal of statistical software* .

Tutz, G., Schmid, M. et al. (2016), *Modeling discrete time-to-event data*.

Vanclay, J. K. (2007), 'On the robustness of the h-index', *Journal of the American Society for information Science and Technology* .

Yang, K. and Meho, L. (2007), 'Citation analysis: A comparison of google scholar, scopus, and web of science', *Proceedings of the American Society for Information Science and Technology* .