Università
Ca'Foscari
Venezia

Corso di laurea Magistrale

in Amministrazione Finanza e Controllo

Tesi di Laurea

Vacanze e Benessere Socioeconomico

Le reti Bayesiane per scoprire le relazioni intrinseche ai movimenti turistici degli Italiani

Relatore
Ch. Prof.ssa Debora Slanzi

Laureando:
Leonardo Rigo
Matricola: 857052

Anno Accademico
2021 / 2022

INDICE

Vacanze e benessere socioeconomico:
Le reti Bayesiane per scoprire le relazioni intrinseche ai movimenti turistici degli Italiani

Capitolo 1
Introduzione

Nel 2010 una collaborazione nata tra Istat, l'Istituto Nazione di Statistica, e CNEL, l'Istituto Nazionale dell'Economia e del Lavoro, ha portato all'avvio del Progetto BES[1], acronimo di Benessere Equo e Sostenibile. Lo scopo di questo progetto, come suggerisce il nome, è quello di misurare il benessere reale dei cittadini italiani attraverso accurate rilevazioni statistiche, il tutto, però, con una determinante consapevolezza di fondo.

Questa consapevolezza riguarda l'inadeguatezza del Pil, Prodotto Interno Lordo, come indicatore di benessere, poiché spesso il carattere puramente economico di questo aggregato tralascia dinamiche sociali e ambientali che possono avere importanti conseguenze sul benessere dell'individuo. Alla luce di ciò sono stati individuati 12 domini rilevanti per il benessere suddivisi in oltre centocinquanta indicatori maggiormente pertinenti a tutta la sfera sociale tralasciata dal Pil per dare reale testimonianza delle condizioni socioeconomiche e, al contempo, creare un consistente e duraturo programma di ricerca.

In questo contesto si colloca questo elaborato, poiché nella strutturazione del BES, manca un'analisi del benessere dal punto di vista turistico-vacanziero, prospettiva forse meno rilevante per il benessere puramente economico ma dai non trascurabili risvolti di equità e socialità.

In questa tesi si farà riferimento agli anni 2014, 2018, 2019 e 2020 per un particolare insieme di microdati rilevati dall'Istat nell'indagine sulle spese delle famiglie al focus "Viaggi e Vacanze" con l'obiettivo di individuare e modellare le dinamiche dei flussi

[1] Si apporta di seguito l'URL per l'ottenimento della versione completa dell'ultimo rapporto BES a disposizione: https://www.istat.it/it/files//2021/03/BES_2020.pdf

vacanzieri per i soggetti residenti nel territorio Italiano in termini di relazioni tra variabili contenute nell'indagine[2].

Lo strumento statistico che consentirà di fornire informazioni circa le relazioni tra le variabili individuate nello studio in oggetto è la rete Bayesiana, una classe di modelli grafici probabilistici che quantifica l'incertezza di un fenomeno attraverso distribuzioni di probabilità associate ai legami tra le variabili del sistema[3]. Specificatamente, attraverso le potenzialità offerte dalle reti Bayesiane, lo scopo di questa ricerca sarà duplice: da un lato fornire agli operatori del settore informazioni chiave riguardanti questo fenomeno al fine di poterlo comprendere a pieno e, interpretando i risultati, per aggiustare l'offerta turistica, quantunque possibile*;* dall'altro, invece, per taluni soggetti istituzionali la rilevanza delle informazioni emerse potrà essere di supporto ad analisi di tipo socio-economico del contesto Italiano.

Lo studio delle "vacanze", infatti, oltre a fornire informazioni chiare e dirette come le località e i periodi dell'anno preferiti dal "viaggiatore", permette, indirettamente, di ricavare informazioni sul benessere, in questo caso, della popolazione Italiana.

Capacità di spesa, titolo di studio, condizione lavorativa sono solo alcune delle variabili che permetteranno all'analisi di fornire un'informazione più precisa e stratificata, nonché l'opportunità di un confronto intertemporale che renderà possibile sia la valutazione di trend sulle relazioni delle variabili d'esame, rispetto ad esempio allo sviluppo tecnologico o al verificarsi di particolari eventi, ma anche di ipotizzare previsioni per il futuro.

La presenza dell'anno 2020 in questo processo di studio inoltre trova motivo di interesse in relazione a quanto l'effetto pandemico del Covid-19 per le sue ricadute a livello socioeconomico abbia messo notevolmente sotto stress il contesto Italiano specialmente in questo settore.

Pertanto, i dati selezionati saranno utilizzati per la creazione di reti Bayesiane, di anno in anno confrontabili, le quali, grazie alle loro proprietà grafico-statistiche, permetteranno un dispiegamento delle informazioni per il finale supporto agli operatori del settore e agli enti istituzionali oltre che per valutare i progressivi eventuali cambiamenti di evidenze nelle relazioni tra le variabili rilevate, anche alla luce di eventi come la pandemia.

[2] Si apporta di seguito l'URL che indirizzerà alla pagina dell'Istat in cui sono messi a disposizione le campionature qui analizzate e le informazioni a loro corredo: https://www.istat.it/it/archivio/178695

[3] Si faccia riferimento alle appendici della tesi per una completa disamina delle variabili presenti all'interno dei dataset

1.1 Evidenze espresse dal BES[4]

In questi dieci anni di vita di questo progetto è stato possibile delineare un quadro abbastanza preciso sulle condizioni della popolazione Italiana; i campi di analisi sono ampi ed eterogenei, molti di questi, infatti, spaziano ben oltre l'orizzonte delle vacanze e il tempo libero, tuttavia, ci si limiterà a presentare brevemente le evidenze ottenute per quei domini più pertinenti all'analisi dei comportamenti vacanzieri e delle conseguenti stratificazioni sociali.

In tema di salute le rilevazioni mostrano un'evoluzione positiva della speranza di vita per l'intera popolazione Italiana, tuttavia, emergono comunque differenze sull'andamento di questa progressione soprattutto in termini principalmente geografici e limitatamente di genere; come si vedrà per altre categorie, essenzialmente, il centro e sud Italia "viaggiano" più lentamente del nord Italia.

Per quanto riguarda, infine, l'impatto della pandemia, esso ha pressoché annullato il guadagno sulle prospettive di vita, però, visto l'unicità di questo evento, è lecito aspettarsi una ripresa del trend interrotto nel 2019, in cui le prospettive di vita medie erano di 82,96 anni.

Parlando di Istruzione l'ultimo decennio vede progressi anche su questo campo, però emerge ancora l'incapacità di offrire agli studenti le medesime possibilità e mezzi, questo denota una forte dipendenza tra istruzione ed estrazione sociale, in cui generalmente le fasce più alte risultano privilegiate, con ciò non viene meno l'effetto di ascensore sociale che determinati titoli di studio possono garantire.

Su scala europea, l'Italia non si presenta al meglio, risulta difficile rispettare la progressione media dei paesi dell'eurozona, indipendentemente dalla fascia di età presa in considerazione.

Per il settore lavorativo, l'effetto della pandemia ha portato conseguenze negative sul tasso di occupazione, che perde due punti percentuali rispetto al 2019, assestandosi al 62%; come per l'istruzione, anche in questo settore i divari con l'Europa sono aumentati, specialmente in relazione alle differenze di genere che evidenziano una maggior penalizzazione per le donne.

[4] Si inserisce l'URL che renderà accessibile una versione ridotta delle evidenze espresse nel BES così da fornire un pronto riscontro sui dati menzionati: https://www.istat.it/it/files//2021/03/BES_2020-nota-stampa.pdf

La retribuzione, dopo anni di andamento stabile, vede nel corso del 2020 un notevole aumento di retribuzioni a paga bassa (una retribuzione oraria inferiore ai due terzi della paga oraria mediana) distribuite maggiormente nel sud Italia, per ridursi salendo lo stivale.

Dopo anni di stabilità nelle ripartizioni, cresce la quota di lavoratori impiegati con contratti a termine per lunghi periodi di tempo, contestualmente è aumentato anche il part-time involontario, stesso avviene anche per la quota dei NEET (giovani non impiegati che non studiano).

In riguardo al benessere economico, la situazione delle famiglie Italiane, che storicamente si distinguono per un alta propensione al risparmio, al vivere in case di proprietà e limitando al massimo il ricorso a forme di debito bancario, vede un lento processo d'uscita dalla crisi dello scorso decennio; sebbene questo periodo difficile, in cui le disuguaglianze e le differenze territoriali si sono fatte più marcate, a partire dal 2018, anche grazie ad una serie di riforme, si è rilevata una serie di miglioramenti, quali: l'incremento dell'occupazione, la riduzione dell'indice di povertà assoluta e l'aumento di reddito e potere d'acquisto per le famiglie, che hanno portato ad un miglioramento nelle condizioni economiche per il paese.

Tutto ciò, però, è stato quasi annullato dall'arrivo del Covid-19, il quale ha riportato l'Italia indietro di molti anni soprattutto in termini di povertà per le fasce più delicate che, all'opposto di quanto ci si aspetterebbe, ha colpito maggiormente le regioni del nord Italia; globalmente dopo la pandemia sono stati rilevati circa un milione e 346mila bambini e ragazzi poveri, 209mila in più del 2019.

Oltre a ciò, è stato rilevato dalle interviste un peggioramento delle condizioni economiche per le famiglie (28,8% degli intervistati dichiara questo peggioramento contro il 25,8% del 2019), con quelle composte da 3 o più componenti a soffrirne maggiormente; si mantengono stabili, invece, le condizioni delle famiglie con gradi di istruzione più elevati o composte prevalentemente da anziani.

Tralasciando il 2020, in ultimo, si evidenziano fino al 2019 cali: del rischio di povertà, attestabile all'epoca al 20,1%, e del tasso di persone in gravi condizioni di deprivazione materiale ed abitativa.

Per quanto concerne, infine, il benessere soggettivo nel 2020 il 44,5%, in aumento rispetto al 2019, della popolazione Italiana esprime un voto tra 8 e 10 per indicare il proprio soddisfacimento sui momenti della vita e sul godimento del tempo libero; la categorie mantiene salde le differenze territoriali tipiche del contesto sociale Italiano, per cui il nord

risulta tendenzialmente privilegiato come anche chi risponde a determinate classi sociali: in relazione a titolo di studio, occupazione e anche genere.
L'isolamento dovuto al Covid-19 e alle conseguenti misure di distanziamento sociale ha avuto ricadute per i gruppi di popolazione che vivono da sole, per le quali cresce la percentuale di insoddisfatti; oltre a ciò, un ulteriore ricaduta negativa della pandemia, vede una riduzione percentuale delle persone che vedono un miglioramento delle proprie condizioni economiche nei prossimi cinque (contestualmente è aumentata la percentuale di chi vede peggiorare le proprie prospettive).
In ogni caso, l'ultima cosa ad emergere è un aumento della percentuale di individui molto o abbastanza soddisfatti del proprio tempo libero, dal 68% del 2019 al 69,4% del 2020, conferma che nonostante tutto vi è una tendenza al miglioramento delle condizioni di vita

1.2 Situazione turistica italiana

Il settore turistico in Italia si caratterizza per l'emersione di alcune problematiche di fondo che ne impoveriscono le potenzialità; sebbene lo scopo di questa analisi non sia quello di risolvere questi problemi, un raffronto con questi sulla base di quanto emergerà dalla stima dei modelli proposti, renderà possibile contestualizzare le informazioni ottenute e giustificare determinate evidenze.
Nello specifico, in un contesto pre-pandemico, i principali problemi che attanagliano il turismo in Italia riguardano: il sovraffollamento di determinate destinazioni, il quale, combinato con la forte stagionalità dei flussi, porta ad un costante overbooking di alcune località.
Questi effetti combinati portano ad una esasperazione delle logiche di mercato causando una costante *race to the bottom* per le offerte proposte, rendendo così il turismo una pura espressione di consumismo, non più orientata ai piaceri della scoperta e del viaggio in sé
Le ripercussioni di questo sfruttamento non si limitano ad impoverire il valore offerto dai territori visitati, ma offuscano anche il panorama delle offerte turistiche meno rispondenti alle precedenti logiche di fruizione.
Di conseguenza il settore vive di poca organizzazione dei flussi al suo interno, i quali nel tempo, aumentando sempre più di dimensione e si limitano a consolidarsi nelle mete e nelle stagionalità portando ad un impoverimento del panorama italiano.
La doppia valenza sia di pura informazione turistica e indirettamente di termometro del benessere sociale del campione statistico selezionato ruoterà attorno allo studio di

numerose variabili chiave che aiuteranno a far comprendere meglio il fenomeno della situazione turistica italiana e ad elaborare considerazioni sulla base delle informazioni estratte.
Le lacune dal punto di vista dello studio del tempo libero presenti nel Bes, il quale si limita a presentare degli indicatori di semplice soddisfazione su come viene passato questo tempo, e la presenza di altre variabili, seppur già inserite nel Bes in altri domini di rilevamento, ci danno qui l'opportunità, invece, di essere analizzate in riferimento al contesto vacanziero cosicché i risultati saranno poi riportati e valutati secondo le più profonde logiche di interpretazione del benessere economico e sociale nonché con fini di supporto al settore.
Regione di provenienza, istruzione, posizione lavorativa e capacità di spesa giornaliera sono solo alcune delle variabili presenti nei dati dell'Istat che saranno in grado di farci dipingere al meglio una rappresentazione del contesto di riferimento.

1.3 Struttura della tesi

Lo scopo di questa tesi è di presentare il modello statistico delle reti Bayesiane per mostarne le potenzialità al fine di rilevare le relazioni tra le variabili che caratterizzano un fenomeno oggetto di studio.
La struttura di questo elaborato è la seguente.
Nel Capitolo 2 si presenterà la teoria delle reti Bayesiane e le informazioni di base per comprendere alcuni concetti statistico-probabilistici utili per la descrizione approfondita delle proprietà del modello e al fine di presentare i risultati d'analisi.
Nel Capitolo 3 sarà introdotto il caso studio preso in esame e i dati su Viaggi e Vacanze disponibili dell'Istat. Verrà presentato il processo di pulizia dei dati campionari e alcune considerazioni su alcune semplificazioni adottate. Una volta fatto ciò si procederà con una descrizione delle variabili chiave selezionate per l'analisi e della loro importanza sull'esito dello studio per poi giungere, finalmente, nel Capitolo 4 all'implementazione delle reti Bayesiane per il fenomeno oggetto di studio. I risultati ottenuti porteranno alla formulazione di alcune conclusioni sugli esiti del processo statistico e si attuerà una comparazione per i vari anni di studio per vedere l'eventuale emersione di trend o macro-relazioni visibili solo olisticamente.
Infine, si presenteranno alcune conclusioni su tutto il processo di elaborazione effettuato.

Capitolo 2
Le reti Bayesiane

Le reti Bayesiane appartengono alla categoria dei network probabilistici: un insieme di modelli grafici in grado di rappresentare i legami di probabilità intercorrenti tra le variabili di un dataset.

Tali modelli si distinguono per l'intuitività dell'informazione espressa, di modo che risulta più facile comprendere e comunicare le relazioni di probabilità condizionata presenti nel dominio di studio; inoltre, questa tipologia di reti fonda le sue proprietà principali nella teoria dei grafi, i quali, nella loro forma più semplice, sono così definiti: *Un grafo si identifica da una coppia* $\mathcal{G} = (V, E)$*, dove:* V *rappresenta un gruppo finito di vertici o nodi distinti, mentre* $E \subseteq V \times V$ *definisce un insieme di archi che esplicitano le connessioni tra i vari nodi; la coppia ordinata* $(u, v) \in E$ *indentifica, invece, la presenza di un arco orientato creante una connessione diretta tra il nodo* u *e il nodo* v*, tale per cui* u *sarà considerato il nodo genitore mentre* v *il rispettivo nodo figlio (figura 2.1).*

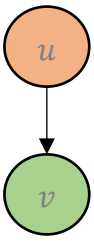


Figura 2.1: Esempio di Grafo

Prima di concentrarsi sui network e le loro proprietà, è bene soffermarsi su un breve riepilogo di alcune nozioni probabilistiche di base, questo perché, avendo a che fare con la determinazione dell'incertezza, la teoria delle reti Bayesiane si fonda sul calcolo probabilistico per esplicitare a pieno le sue funzionalità.

2.1 Cenni alla teoria della probabilità

Storicamente il calcolo della probabilità fu inventato nel diciassettesimo secolo da Fermat e Pascal dall'esigenza pratica di fronteggiare l'incertezza intrinseca al gioco d'azzardo; visti i risvolti ottenuti in questo contesto, tuttavia, non ci volle molto perché questa nuova scienza fosse applicata in altri campi di interesse.

Uno dei modi più semplici per affrontare questa materia e renderla di conseguenza anche più chiara e intellegibile, avviene mediante la trasposizione visiva del fenomeno che si è intenti a studiare; questo modus operandi traccerà, poi, una sorta di filo rosso all'interno di questa narrazione, testimonianza che nonostante le complessità assunta da determinati problemi la capacità di saperli modellizzare in supporti visivamente accessibili risulta tuttora una soluzione vincente.
Partendo, quindi, dalle basi, i diagrammi di Venn (si veda un esempio in Figura 2.2) aiutano a comprendere e rendere più chiare le dinamiche del calcolo probabilistico, le quali, successivamente, fungeranno da fondamenta per tutte le applicazioni più complesse che si andranno ad affrontare.
Prima di parlare di probabilità vera e propria, però, è bene esaminare lo svolgersi di un fenomeno al fine di attribuire il corretto significato alle parti che lo compongono; detto ciò, per l'analisi dei fenomeni aleatori è necessario, in primis, stabilire il corretto dominio di studio, di conseguenza, si andranno qui a definire: "esperimenti" tutti quegli eventi governati dal caso e l'aleatorietà, mentre "spazio campionario" indicherà l'insieme di tutti gli esiti possibili di questi eventi.

Esempio lancio di una moneta

Si suppone che l'esperimento veda il lancio di una moneta, ad esito di tale processo si avranno due alternative finali: che la faccia rivolta verso l'alto sia testa (T) oppure croce (C), perciò sulla base di quanto detto poc'anzi si può indicare lo spazio campionario (Ω) come:

$$\Omega = \{T, C\}$$

Oltre a ciò, è importante sapere che uno spazio campionario (Ω) può essere di rilevante importanza per il ruolo dei sottoinsiemi che lo compongono

Esempio dado a sei facce

Sapendo che lo spazio campionario di un dado a sei facce comprende tutti gli esiti possibili ottenibili da un suo lancio perciò:

$$\Omega = \{1,2,3,4,5,6\}$$

Sulla base di questo si possono supporre determinate eventualità o eventi (E) tali per cui si potranno identificare dei sottoinsiemi di Omega:

-facce pari $Ep = \{2,4,6\}$
-facce dispari $Ed = \{1,3,5\}$

Sulla base proprio di questi eventi, infine, si andranno poi a calcolare le probabilità dei loro accadimenti.

Per una più generica definizione ad ampio spettro si può identificare la probabilità come: data la presenza di un fenomeno casuale o aleatorio, E, la probabilità dell'accadimento di tale evento, $P(E)$ dove $E \in \Omega$, si esprime con un numero compreso fra 0 e 1, che permette di capire il grado di eventualità che l'evento si verifichi, intendendo che il valore minimo 0 corrisponda al caso in cui l'evento sia impossibile, mentre il valore massimo 1 corrisponda al caso in cui l'evento sia certo.

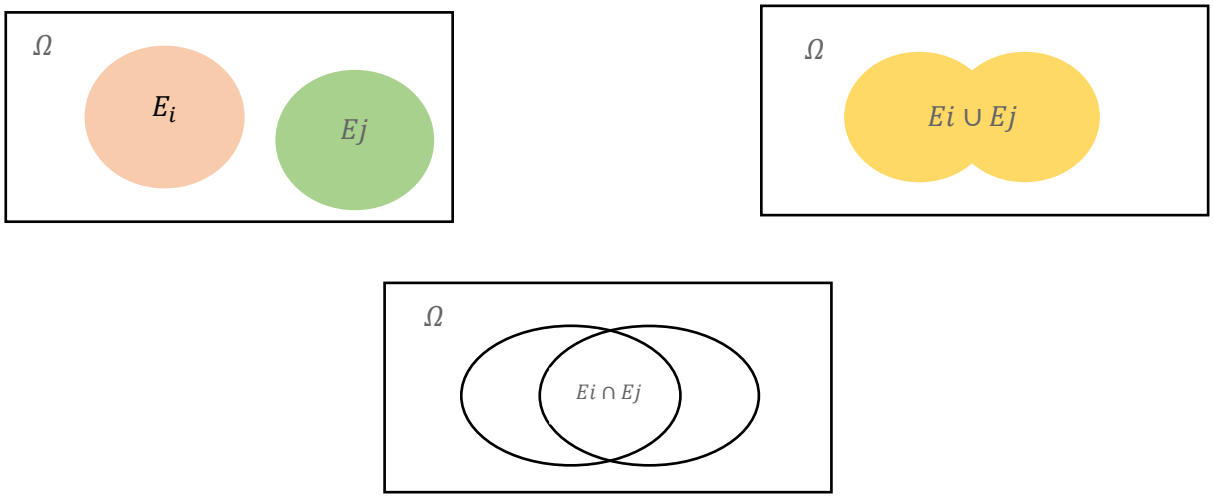


Figura 2.2: Diagrammi di Venn

2.1.1 Assiomi della probabilità

Con alla base questa definizione iniziale e grazie agli studi del matematico russo A.N. Kolmogorov è stato poi possibile identificare degli assiomi che regolano le dinamiche del calcolo probabilistico.

- La probabilità di un dato evento E è sempre compresa tra 0 e 1 e non può mai essere negativa

$$0 \leq P[E] \leq 1$$

- Per l'intero spazio campionario Ω vale la seguente relazione

$$P[\Omega] = 1$$

- Presa una successione di eventi $E1, E2, En$ ... tali che $Ei \cap En = \emptyset$ con $i \neq n$

$$P[Ei \cup En] = \sum P[Ei]$$

- La probabilità di un insieme vuoto è sempre pari a zero

$$P[\emptyset] = 0$$

- Risulta possibile calcolare la probabilità di un dato evento a partire dalla probabilità che ha questo di non realizzarsi

$$P[E] = 1 - P[\bar{E}]$$

- Presi due eventi Ei e Ej la probabilità della loro unione viene calcolata come la somma delle probabilità dei singoli eventi al netto delle loro intersezioni

$$P[Ei \cup Ej] = P[Ei] + P[Ej] - P[Ei \cap Ej]$$

- Qualora Ei e Ej risultassero tra loro come eventi indipendenti si avrebbe che la loro intersezione sarebbe uguale a zero, $P[Ei \cap Ej] = 0$, perciò la loro unione risulterebbe solo:

$$P[Ei \cup Ej] = P[Ei] + P[Ej]$$

2.1.2 Definizione di Probabilità

Esistono diverse versioni della definizione di probabilità, a seconda della teoria che si considera.
Nella teoria classica si definisce probabilità di un evento E il rapporto tra i casi favorevoli all'avverarsi di tale evento sul totale di casi possibili purché l'insieme dei casi possibili sia un insieme finito e ogni caso sia equiprobabile.

Probabilità di $E = casi\ favorevoli/casi\ possibili$

Esempio estrazione da un'urna

Supponiamo che un'urna contiene 5 biglie, 2 blu e 3 rosse; quindi, la probabilità di estrarre una biglia rossa sarà il risultato del seguente rapporto in cui biglia rossa sarà il caso favorevole e l'insieme delle biglie nell'urna i casi possibili

$$Pr = 3\ biglie\ rosse/5\ totale\ delle\ biglie$$

La definizione frequentista, a differenza di quella classica, basa i suoi costrutti sulla natura riproducibile di un dato esperimento in analoghe situazioni che mantengono inalterate le sue condizioni; perciò, si definisce la probabilità di un evento E il rapporto tra il numero di successi nell'esperimento e il numero totale di prove.
Un buon esempio di ciò è il semplice lancio della moneta, un esperimento replicabile e da precisi esiti.

$$Probabilità\ di\ E = numero\ di\ successi/numero\ di\ prove$$

Un altro aspetto chiave da conoscere per i suoi usi e applicazioni è quello riguardante la probabilità condizionata, questo tipo di probabilità mette in relazione due eventi, che chiameremo E ed F, di modo che sia possibili calcolare la probabilità connessa all'accadimento dell'evento E data però anche la condizione dell'accadimento dell'evento F.
È necessario introdurre il concetto di dipendenza tra eventi:

Eventi dipendenti: si definiscono dipendenti due eventi compatibili E e F qualora al verificarsi di uno dei due vi sia un'alterazione alla probabilità di verificarsi dell'altro, un esempio di ciò è l'estrazione senza reinserimento di biglie da un'urna in cui ogni estrazione altererà le probabilità delle successive.

Pertanto, sarà possibile calcolare la probabilità condizionata mediante la seguente formula:

$$P[E|F] = (P[E \cap F])/P[F]$$

Da cui si ricaveranno poi le seguenti relazioni:

$$P[E \cap F] = P[E|F]P[F] \quad e \quad P[E \cap F] = P[F|E]P[E]$$

Dalle equazioni mostrate sopra si evince quanto detto, ovvero che: la probabilità dell'evento E dipende dal verificarsi dell'evento F, evento che risponde anch'esso ai precedenti assiomi probabilistici.
In una prospettiva grafica, riconducendo a quanto già fatto con Venn per la probabilità congiunta possiamo vedere la probabilità condizionata di E su F come l'intersezione dei due eventi sul totale evento F.
Dal concetto di dipendenza si giunge infine a quello di indipendenza che per simmetria identifica la probabilità condizionata di E e F, qualora questi siano in uno stato di indipendenza.
Gli eventi indipendenti si caratterizzano per la mancanza di reciproche ripercussioni al verificarsi di uno dei due eventi, comportando, sulla base delle medesime relazioni della probabilità condizionata, che:

$$P[E|F] = P[E] \quad e \quad P[F|E] = P[F]$$

Sulla base di questo, risulta poi ricavabile:

$$P[E \cap F] = P[E]P[F]$$

2.1.3 Teorema di Bayes

Un ulteriore formula che è stata introdotta nel campo della probabilità condizionata è attribuita al matematico inglese Thomas Bayes che nel 1700 notò la possibilità di calcolare $P[E|F]$ partendo da $P[F|E]$.
Concettualmente questa inversione permette di considerare F come accaduto e capire la probabilità connessa al fatto che F si sia verificato proprio per effetto dell'evento E, perciò sulla base di questo è stato formulato che:

Dati l'evento F e una partizione $\{Ei\}$ di Ω si ha che ∀ h:

$$P[Eh|F] = P[F|Eh]P[Eh]/\sum i\ P[F|Ei]P[Ei]$$

Per essere più precisi, il teorema di Bayes permette di dire che la probabilità di un'ipotesi h condizionata ad un evento e è uguale al suo inverso; quindi, la probabilità dell'evento e condizionato all'ipotesi h, moltiplicato per la probabilità della sua condizione a priori p(h), il tutto diviso per la probabilità dell'evento e.
Un ulteriore ragionamento sulla formula di Bayes rende possibili grazie all'assioma $P(A\ e\ B) = P(A,B) = P(B|A)P(A) = P(A|B)P(B)$ e alla regola della probabilità totale si può riscrivere la formula di Bayes applicando la seguente semplificazione:

$$p(h|e) = p(e|h)p(h)$$

Rendendo difatti direttamente proporzionali la probabilità dell'ipotesi h dato l'evento e per il prodotto tra il suo "inverso" e la nuda probabilità di h.

2.2 Le reti Bayesiane

Supponiamo di avere tre variabili (X_1, X_2, X_3) e che ognuna di esse mostra delle influenze sulla successiva, ovvero: X_1 condiziona X_2 che a sua volta condiziona X_3 ma non ci sono relazioni dirette tra X_1 e X_3; la rappresentazione grafica di questo diagramma di influenze può essere ottenuta con una struttura molto semplice di nodi e frecce, come si può vedere

nella Figura 2.3, mentre risulta leggermente più sofisticata la formulazione di un equazione in grado di descrivere questo fenomeno relazionale.

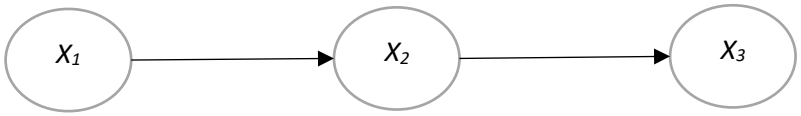


Figura 2.3 Diagramma di propagazione delle influenze su tre variabili in serie

$$P(X_1, X_2, X_3) = P\big(X_1\big|Pa(X_1)\big)P\big(X_2\big|Pa(X_2)\big)P(X_3|Pa(X_3))$$

L'equazione sovrastante rappresenta, in una casistica a tre variabili come quella che si sta valutando, la definizione formale di una rete Bayesiana e rende possibile convertire la probabilità di X_1, X_2, X_3 come il prodotto delle rispettive probabilità condizionate al nodo genitore indicato con: $Pa(X_i)$.

Applicando il medesimo processo logico possiamo estendere la modellazione al caso in cui vengono prese in considerazione n variabili casuali $(X_1, X_2, \dots X_n)$, per cui risulta possibile individuare un grafo con n nodi, in cui ogni nodo si associa con la rispettiva variabile (il nodo $j[1, n]$ sarà quindi associato alla variabile X_j) allora il grafo sarà una rete Bayesiana se la rappresentazione delle variabili $X_1, X_2, \dots X_n$ rispecchiasse la seguente equazione:

$$P(X_1, X_2, \dots X_n) = \prod_{j=1}^{n} P\left(X_j\middle|Pa\big(X_j\big)\right)$$

dove $\mathrm{Pa}\big(\mathrm{X_j}\big)$ indica il gruppo di variabili genitore $\mathrm{X_i}$, tali che descrivono un arco diretto tra l'i-esimo nodo verso il j-esimo nodo figlio.

In altri termini, quindi, una rete Bayesiana può anche essere descritta come un grafo diretto aciclico, o più brevemente "*Dag"*, che definisce la fattorizzazione di una

distribuzione di probabilità congiunta sulle variabili rappresentate dai nodi della "*Dag"*; specificando che tale fattorizzazione è data dagli archi diretti presenti nella "*Dag"*.

La stretta correlazione tra il piano grafico e quello più formale di modellazione statistica rendono le reti Bayesiane uno strumento in grado di portare ad una semplificazione di facile intuizione il complesso di relazioni emergenti dai legami nascosti all'interno dell''apparato statistico oggetto di studio.

"Probability calculus allows us to represent the independencies which other systems require, but also allows us to represent any dependencies which we may need".

Le reti Bayesiane, perciò, attraverso le proprietà della loro struttura grafica permettono di rappresentare e studiare fenomeni dominati dall'incertezza.
I nodi di cui sono composte individuano le variabili del dominio di studio, i quali, a loro volta, vengono collegati e messi in relazione tra loro grazie ad una serie di archi diretti che sottendono alla presenza di relazioni di dipendenza diretta tra le variabili connesse; tali legami si creano grazie alle proprietà della probabilità condizionata applicata e associata ricorsivamente per ogni nodo della rete.
Non esistono particolare dettami sul comportamento e la disposizione dei nodi e delle rispettive relazioni purché non ci siano cicli diretti: ciò implica che il formarsi di una rete non prevede mai la possibilità di ritornare al nodo di partenza seguendo il diramarsi degli archi diretti, in sostanza non vi è spazio per la ricorsività nella rete e per questo motivo questi network vengono anche chiamati *"directed acyclic graphs"* o più semplicemente *"Dags"*, com'è stato preannunciato nei capoversi precedenti.

2.2.1 Nodi e valori

Le informazioni che un nodo è in grado di fornire sono di molteplici tipologie, ma in questo momento ci limiteremo a considerare il caso in cui essi restituiscono solo ed esclusivamente la rappresentazione di variabili categoriali o discrete.
Tale natura discreta del nodo si subspecifica ulteriormente in alcune categorie di uso consueto che potremmo qui definire come:

- *Nodi Booleani che garantiscono informazioni di natura "proposal" come, per esempio, la risposta Vero o Falso ad un determinato quesito*

- *Nodi che rappresentano valori di ordinamento come possono essere le valutazioni sulla difficoltà di tracciato sciistico (facile, di media difficoltà, arduo)*

- *Nodi a valori interi che restituiscono l'informazione in modo diretto un esempio di ciò è l'indicazione di un possibile parametro età che attraverso il nodo restituisce il valore effettivo del soggetto esaminato.*

2.2.2 Struttura

Le regole intrinseche nella definizione della struttura di una rete Bayesiana hanno come principio il far emergere tutte le relazioni qualitative presenti e a volte nascoste tra le variabili; la forma di configurazione più semplice, ma anche punto di partenza per l'elaborazione di strutture più complesse, è quella che prevede la connessione diretta che segue la direzione del flusso "influenzale" tra due nodi qualora uno di questi abbia effetto sull'altro.
Il verso e la direzione dei legami emersi tra i nodi rendono possibile l'utilizzo di una terminologia metaforica per descrivere il grafo e il verso delle influenze che intercorrono in esso: si parlerà quindi di nodi genitori e nodi figli qualora le connessioni siano dirette o ad un livello più macroscopico di nodi "antenati" e nodi "discendenti" qualora la catena di connessioni che congiunge i due nodi sia più lunga, impedendo, perciò, un collegamento diretto.
Un ulteriore concetto topologico di rilevanza è il cosiddetto di *"Markov blanket" (Figura 2.4)*, con questo termine dal punto di vista grafico si intende tutta quell'area spaziale di relazioni limitrofe ad un nodo di riferimento che comprendono i suoi nodi genitori, i suoi nodi figli e i loro eventuali ulteriori genitori.

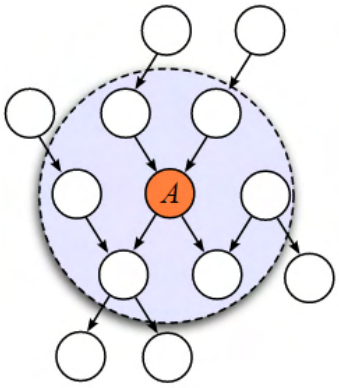


Figura 2.4: Esplicitazione visiva dell'area spaziale rispondente alla definizione di "Markov blanket"

Per descrivere la struttura di queste reti ci si rifà alle analogie tra queste e la forma degli alberi in modo tale che i nodi privi di genitori siano chiamati nodi *"radice",* quelli privi di figli siano, invece, chiamati nodi *"foglia"*, mentre per i nodi che non rispondono alle proprietà di queste due categorie vi è la semplice nomenclatura di nodo *"intermedio"* a garanzia della loro identificazione.

È bene sottolineare che: il punto di vista convenzionale per un esame visivo della rete prevede che questa si dirami dall'alto verso il basso, quindi, in riferimento alla struttura ad albero, essa viene perciò rappresentata a rovescio rispetto a quanto ci si aspetterebbe: con le radici in alto e le foglie in basso.

La modellazione di una rete Bayesiana sottende, in generale, al rispetto della suddetta proprietà di Markov, la quale predispone che: "non ci sono legami di dipendenza diretti nel modello di riferimento che non siano già stati esplicitati attraverso archi tra i vari nodi"; di conseguenza le reti Bayesiane che rispettano questa proprietà si figurano come *"mappe di indipendenza"* (*I-maps*), quando la mancanza di archi suggerisce una reale mancanza di dipendenza nel sistema.

Discorso analogo segue qualora vi sia la presenza di archi, i quali, sebbene a volte non sottendono ad una reale rapporto di dipendenza nel sistema, se corrispondono a dipendenza diretta fanno sì che il network si possa definire una *"mappa di dipendenza"* (*D-maps*); la presenza di *I-maps,* tuttavia, non esclude la presenza di *D-maps* e viceversa,

ma, anzi, qualora si trovino compresenti all'interno della stessa rete fanno sì che essa si possa definire come una *"mappa perfetta"*.[5]

2.2.3 Indipendenza condizionale

Le reti Bayesiane che soddisfano la proprietà di Markov sono in grado di esprimere relazioni di indipendenza condizionale attraverso la distribuzione di probabilità, le quali sono, a loro volta, capaci di modificare i costrutti informativi sulle cause nonché di influenzare la comprensione della struttura in sé.

Catene seriali

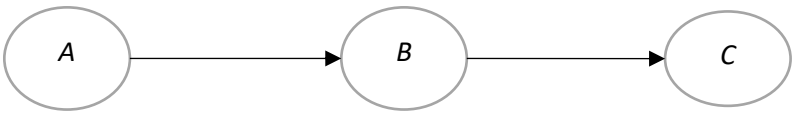


Figura 2.5: Esempio di catena seriale a tre variabili

$$P(C|A \wedge B) = P(C|B)$$

Si è in presenza di una catena seriale, qualora, considerando tre distinti nodi A, B e C in progressiva relazione diretta tra loro si ha che: la probabilità di C dato B è la medesima di C dati sia A che B, questo perché A non ha alcun effetto sulle evidenze espresse da C se sappiamo che B è già accaduto; è possibile riscrivere quanto detto attraverso questa formulazione: $A \perp C|B$, la quale ci fa comprendere meglio come si sviluppa il flusso informativo nel diagramma, poiché la conoscenza di B impedisce lo scambio di informazioni tra A e C e viceversa.

[5] Si rimanda alla lettura del capitolo 2.2.2 del libro: Korb & Nicholson (2004), *Bayesian Artificial Intelligence*, Chapman & Hall/CRC, Boca Raton, Florida

Cause comuni

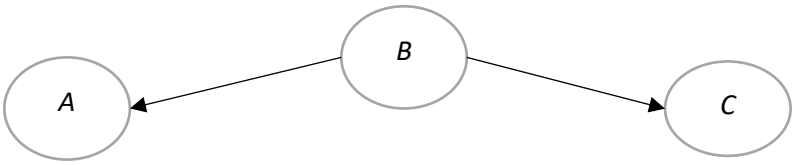


Figura 2.6: Esempio di propagazione di una causa comune B sui rispettivi nodi figlio

$$P(C|A \wedge B) = P(C|B) \equiv A \perp C|B$$

Si verifica uno scenario di casualità comune quando, presi sempre i tre nodi A, B e C, si ha che: A e C dipendono da B in quanto nodo genitore dei primi due, anche in questo caso come nel precedente B impedisce lo scambio di informazioni tra A e C e viceversa.

Effetti comuni

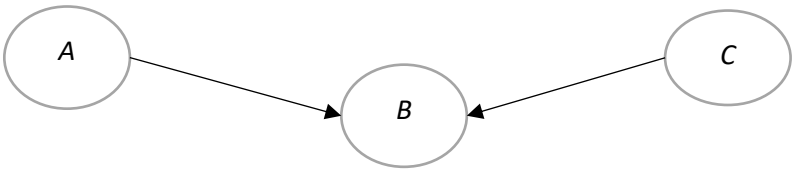


Figura 2.7: Esempio di effetti comuni data l'indipendenza dei nodi A e C

$$P(A|C \wedge B) \neq P(A|C) \equiv \neg(A \perp C|B)$$

Questa tipologia di caso si distingue per la caratteristica struttura a V e rappresenta l'eventualità in cui un singolo nodo, B, ha all'origine due diversi nodi, A e C.
La presenza di effetti comuni è l'esatto opposto di quanto si produce mediante l'indipendenza condizionale nei due casi precedenti, inoltre, sebbene A e C appaiano

indipendenti le informazioni riguardanti l'effetto comune, B, fa sì che essi entrino in rapporto di dipendenza.

d-separation

I concetti appena descritti trovano applicazione non solo in situazioni in cui sono presenti pochi nodi ma anche all'eventualità di una notevole molteplicità di variabili; in vero, grazie all'applicazione della proprietà di Markov: *"in una reete Bayesiana, presa una variabile e dati i relativi nodi genitore, essa sarà indipendente da tutti i suoi nodi non figlio";* è possibile determinare qualora un set di nodi X sia o meno indipendente da un set di nodi Y date le informazioni e le evidenze su un set E, per fare ciò è imprescindibile introdurre alcune definizioni nonché il concetto di d-separation (il quale viene presentato visivamente in figura 2.8),

1. Sentiero (path): dati due set di nodi X e Y, con questo termine si identifica una serie di relazioni che si sviluppa tra un membro di X e un membro di Y tali che ogni nodo è connesso con i suoi adiacenti (indipendentemente la direzione) e nella sequenza non vi si riscontra nessuna apparizione plurima del medesimo nodo.

2. Sentiero bloccato: un sentiero si dice bloccato quando, dato un set di nodi E, ci si trova in presenza di un nodo Z che presenta una delle seguenti condizioni:
    a. Z è in E e Z presenta una struttura degli archi a catena seriale
    b. Z è in E e Z presenta una struttura degli archi a cause comuni
    c. Z e neppure un suo discendente appaiono in E ed entrambi gli archi conducono a Z mediante una struttura ad effetti comuni

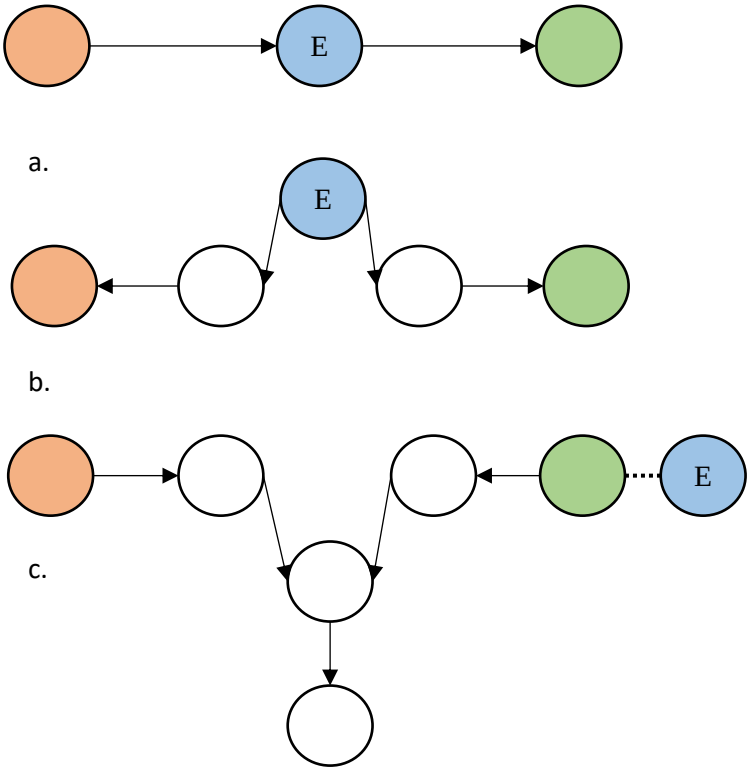


Figura 2.8: Esempi di sentieri bloccati dall'evidenza E per cui i nodi in rosa e in verde possono considerarsi d-separati

Si definisce dunque il concetto di d-separation: dato un set di nodi E questo d-separa altri due set di nodi X e Y se un qualsiasi sentiero da un nodo in X a un nodo in Y risultasse bloccato da un nodo E.
Ciò comporta che: nel caso X e Y apparissero d-separati da E, essi sarebbero tra loro in una condizione di indipendenza condizionale secondo la proprietà di Markov.[6]

2.2.3 Distribuzioni di probabilità condizionata: i parametri della rete

Una volta introdotti i concetti fondamentali che guidano la descrizione e la strutturazione del network è essenziale portare il focus dell'attenzione sulla natura delle relazioni che si sviluppano tra i nodi; per fare ciò risulta necessario stabilire le distribuzioni di probabilità condizionata presenti tra ogni singolo nodo e, poiché si stanno considerando per il

[6] Si rimanda alla lettura del capitolo 2.4.4 del libro: Korb & Nicholson (2004), *Bayesian Artificial Intelligence*, Chapman & Hall/CRC, Boca Raton, Florida

momento solo variabili discrete, il risultato di questo processo vedrà la creazione di una tavola di probabilità condizionata (anche detta *"CPT"*).
Dato un nodo, quindi, il punto di partenza prevede l'osservazione di tutte le combinazioni di valori che possono presentare i nodi genitori; ognuna di queste combinazioni viene poi definita come *istanziazione* dell'insieme genitoriale. Successivamente, per ogni istanziazione si andrà a stabilire la probabilità che il nodo figlio mostri determinati valori.
Il medesimo processo per la stesura del CPT si verifica anche per i nodi radice; tuttavia, la natura apicale di questo nodo rende possibile individuare solo la sua probabilità a priori, discorso diverso per i nodi intermedi che qualora abbiano un vasto numero di nodi genitori o quest'ultimi presentano un ampio spettro di valori possibili le dimensioni del CPT potrebbero risultare estremamente ampie.

2.3. Ragionare con le reti Bayesiane: inferenza probabilistica

Stabiliti modello di riferimento e natura delle relazioni di dipendenza e indipendenza condizionata sottostanti ad esso, lo scopo ultimo di una rete Bayesiana è quello di essere utilizzata come strumento per lo sviluppo di un processo di supporto alle decisioni.
Nello specifico, fissando a probabilità certa i valori presentati da alcune variabili è di interesse valutare come si modifica la probabilità degli stati assunti delle altre variabili o di un sottoinsieme rilevante di esse sulla base di queste nuove informazioni.
Questo processo, che deriva dall'apporto di un condizionamento, viene chiamato inferenza probabilistica o *belief updating*, e dirama le sue conseguenze a tutto il flusso informativo del network, anche non seguendo le direzioni imposte dagli archi.
Sapendo, quindi, che la rete dà piena rappresentazione delle relazioni sottostanti ai rapporti delle variabili, questo implica la possibilità che esse possano essere condizionate ad un qualunque grado di parentela.
Sulla base di quanto detto si possono, quindi, identificare almeno quattro diagrammi di ragionamento (Figura 2.9):

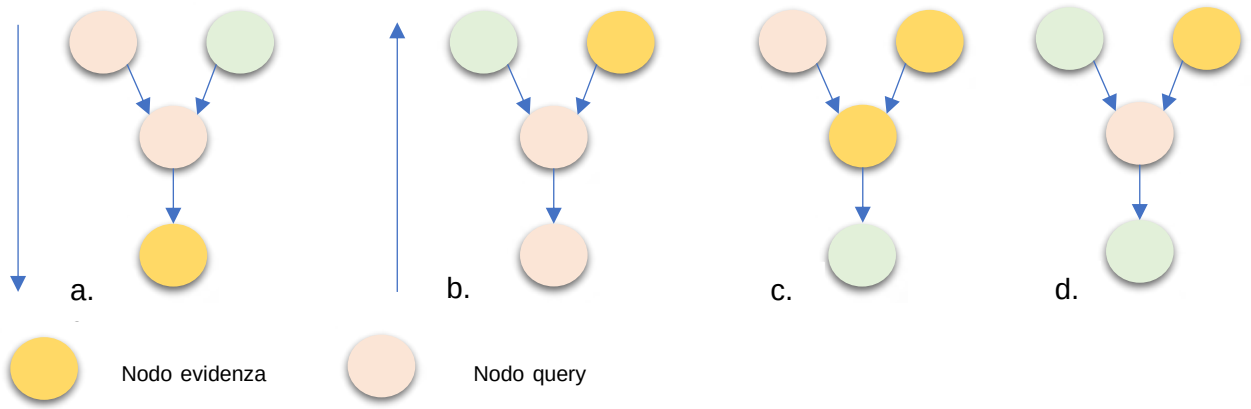


Figura 2.9: Diagrammi di sviluppo del processo inferenziale

a. Diagnostico: permette di ragionare sulle cause partendo dagli effetti indipendentemente dal fatto che questo processo sia in direzione contraria al flusso arcale

b. Predittivo: un ragionamento che si sviluppa dal condizionamento di un nodo antenato grazie all'apporto di nuove informazioni sulle cause che rendono possibile predisporre delle previsioni sui futuri effetti in una logica discendente concorde al verso degli archi.

c. Intercausale: riguarda il ragionamento che possono avere mutue cause sul medesimo effetto comune, un esempio di questo è il caso *"explaining away"* che in presenza di due possibili cause per un unico effetto può essere rappresentato con una struttura a *V* all'interno di una rete Bayesiana.

d. Combinato: questa tipologia, infine, prevede l'utilizzo di tutte le tecniche sovrastanti poiché all'approccio pratico sono pochi i casi in cui prevale un'unica tipologia di ragionamento sulle altre, bensì, spesso risulta necessario l'utilizzo combinato di queste.

Una delle più chiare proprietà delle reti Bayesiane riguarda la possibilità, dunque, di essere utilizzate per ricalcolare le evidenze (o *belief*) qualora nuove informazioni (prove o *evidence*) dovessero presentarsi nel contesto d'analisi.

Questo tipo di prove si differenziano a loro volta in base alla natura qualitativa dell'informazione che sono in grado di fornire, essenzialmente, quindi, si andranno a definire: prove specifiche tutte quelle prove che restituiscono un'informazione precisa su un determinato nodo ($X = x$), mentre, si diranno prove negative quelle che ritornano un'informazione di carattere esclusivo sui possibili valori di un nodo (se $Y = y_1 \, o \, y_2$, evidence: $Y \neq y_1$).

Oltre alle due categorie descritte all'interno delle reti Bayesiane trovano spazio anche le prove *"virtuali",* che non fondano il proprio potere informativo su evidenze comprovate, ma bensì si istaurano su una logica di verosimiglianza e possibilità.

Ricordando ancora una volta che: l''obiettivo primario, per un qualsiasi sistema di inferenza probabilistica, è il saper determinare la distribuzione di probabilità di un gruppo di nodi causali date alcune evidenze sui nodi effetto; nella disciplpina Bayesiana i processi di *"belief updating"* si distinguono ulteriormente in relazione a valutazioni quantitative e qualitative sulle evidenze di cui si è a disposizione.[7]

2.3.1. Inferenza esatta

Catena a due nodi

Il caso più semplice da porre in esame è quello di un network composto esclusivamente da due nodi, $X \longrightarrow Y$.

In questo scenario, al verificarsi di una informazione certa sullo stato de nodo genitore, ad esempio: $X = x$, la probabilità a posteriori sul nodo Y, $Bel(Y)$, può essere desunta direttamente dal valore presente nel *CPT*, $P(Y|X = x)$.

Qualora, invece, vi fossero mutamenti delle evidenze nel nodo figlio, ad esempio: $Y = y$, l'inferenza sul nodo genitore X si otterrebbe mediante l'applicazione del teorema di Bayes:

[7] Per una lettura più approfondita sull'argomento si rimanda alla visione del capitolo 3 del libro: Korb & Nicholson (2004), *Bayesian Artificial Intelligence*, Chapman & Hall/CRC, Boca Raton, Florida

$$Bel(X = x) = P(X = x|Y = y)$$

$$= \frac{P(Y = y|X = x)P(X = x)}{P(Y = y)}$$

$$= \alpha P(x)\lambda(x)$$

Dove:

$\alpha = \frac{1}{P(Y=y)}$, $P(x)$ è la probabilità a priori, mentre $\lambda(x) = P(Y = y|X = x)$ è una stima dell'eventualità.

Si noti che non è necessario conoscere la probabilità a priori dell'effetto; dato che tutti i possibili valori della causa X devono avere somma uno; mediante, poi, la conseguenza del teorema della probabilità totale, è possibile introdurre alfa come costante di normalizzazione.

Catena a tre nodi

Esattamente come fatto per il caso precedente, si mantiene il medesimo iter procedimentale anche se ci si trova di fronte ad una catena in cui vi si riscontra la presenza di tre nodi: $X, Y\ e\ Z$ in relazione progressiva.
In questa casistica evidenze accorse al noto X, $X = x$, si propagano nella rete seguendo la direzione degli archi mediante la logica proposta della regola della catena.

$$Bel(Z) = P(Z|X = x) = \sum_{Y=y} P(Z|Y)P(Y|X = x)$$

Qualora, invece, vi fossero mutamenti delle evidenze nel nodo foglia, $Z = z$, l'inferenza sul nodo genitore X per acquisire $Bel(x)$ si otterrebbe mediante l'applicazione del teorema di Bayes.

$$Bel(X = x) = P(X = x|Z = z)$$

$$= \frac{P(Z = z|X = x)P(X = x)}{P(Z = z)}$$

$$= \frac{\sum_{Y=y} P(Z = z|Y = y, X = x)P(Y = y|X = x)P(X = x)}{P(Z = z)}$$

$$= \alpha P(x)\lambda(x)$$

Dove:

$$\lambda(x) = P(Z = z|X = x) = \sum_{Y=y} P(Z = z|Y = y)P(Y = y|X = x)$$

Polytrees (polialbero o foreste)

La prima tipologia di struttura articolata su cui si può operare un processo inferenziale è la cosidettà polytree; essa si caratterizza per la presenza di al massimo un sentiero che congiunge una qualsiasi coppia di nodi, ciò rende di modo queste reti solo "parzialmente" connesse.

Da qui come mostrato in figura 2.10 è possibile vedere il nodo X espresso nel suo generico polytree dove appaiono chiare le connessioni con i nodi genitori nonché quelle con i nodi figli che a loro volta evidenziano legami con ulteriori genitori formando sub strutture a V; tutto ciò aiuta ad intuire il passaggio locale delle informazioni qualora vi fosse il sopraggiungere di nuove evidenze o mutamenti nelle cause, inoltre, la struttura in sé permette di operare considerazioni diagnostiche o predittive.

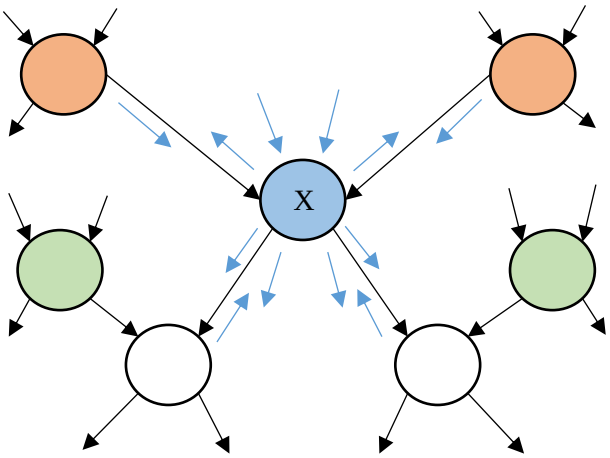


Figura 2.10: Rappresentazione di un generico polialbero in cui le frecce in azzurro rappresentano la propagazione e l'incorporazione delle evidenze da e verso il nodo X

Kim e Pearl: algoritmi di passaggio informazioni

Basato sulla ripetuta applicazione del teorema di Bayes e l'utilizzo delle proprietà di indipendenza condizionata proprie dei network Bayesiani, l'algoritmo si fonda sull'idea che ogni sua interazione porta ad un locale aggiustamento di $Bel(x)$ mediante l'utilizzo di tre parametri: $\lambda(x)$, $\pi(x)$ e le equazioni interne al *CPT*; nello specifico $\lambda(x)$ e $\pi(x)$ sono determinati mediante l'utilizzo delle informazioni π e λ provenienti dai genitori del nodo X e dai suoi figli; le informazioni ottenute attraverso π e λ si propagano, poi, anche ai nodi limitrofi affinché si possano sviluppare eventuali aggiustamenti.[8]

Algoritmo:

L'algoritmo richiede che siano mantenuti i tre seguenti tipi di parametri:

- *ogni connessione di* $U_j \longrightarrow X$ *contribuisce a creare e rinforzare il supporto predittivo di:*

$$\pi_X(U_i) = P(U_i|E_{U_i \setminus X})$$

[8] Per una disamina più dettagliata si prenda visione del capitolo 3.3.1 del libro: Korb & Nicholson (2004), *Bayesian Artificial Intelligence*, Chapman & Hall/CRC, Boca Raton, Florida

- *ogni informazione scaturita da $X \longrightarrow Y_j$ contribuisce creare e rinforzare il supporto diagnostico di λ:*

$$\lambda_{Y_j}(X) = P(E_{Y_j \setminus X}|X)$$

- *si fanno fisse le relazioni probabilistiche all'interno del CPT*

$$P(X|U_i, \dots, U_n)$$

Questi parametri si utilizzano per i processi di aggiornamento locale delle evidenze informative mediante un processo a tre fasi:

1. ***Belief updating***

L'aggiornamento delle evidenze per il nodo X avviene mediante l'apporto di nuovo materiale informativo proveniente dai nodi genitori oppure dai nodi figlio, portando conseguentemente ad un aggiustamento dei parametri connessi alle evidenze.

$$Bel(x_i) = \alpha\lambda(x_i)\pi(x_i)$$

Dove:

$$\lambda(x_i) = \begin{cases} 1 & se\ l'evidenza\ risulta\ X = x_i \\ 0 & se\ l'evidenzariguarda\ un\ altro\ x_i \\ \prod_j \lambda_{Y_j}(x_i) & per\ il\ casi\ rimanenti \end{cases}$$

$$\pi(x_i) = \sum_{u_1,\dots,u_n} P(x_i|u_1,\dots,u_n) \prod_i \pi_X(u_i)$$

Con α come costante di normalizzazione per cui: $\sum_{x_i} Bel(X = x_i) = 1$

2. Propagazione bottom-up

Si riscontra questo tipo di propagazione quando dal nodo X emerge una nuova informazione λ da condividere con i relativi nodi genitore.

$$\lambda_X(u_i) = \sum_{x_i} \lambda\,(x_i) \sum_{u_k:k\neq i} P\,(x_i|u_1,\dots,u_n) \prod_{k\neq i} \pi_X(u_k)$$

3. Propagazione top-down

Di logica diametralmente opposta alla precedente, il flusso informativo vede il propagarsi dell'evidenza dal flusso X in direzione della sua "prole".

$$\pi_{Y_j}(x_i) = \begin{cases} 1 & se\ risulta\ inserito\ x_i\ come\ valore\ di\ evidenza \\ 0 & se\ l'evidenza riuarda\ un\ valore\ x_j \\ \alpha\left[\prod_{k\neq j} \lambda_{Y_k}(x_i)\right] \sum_{u_1,\dots,u_n} P(x_i|u_1,\dots,u_n) \prod_i \pi_X\,(u_i) = \dfrac{\alpha Bel(x_i)}{\lambda_{Y_j}(x_i)} \end{cases}$$

L'equazione qui presentata permette di definire il parametro $\lambda(x_i)$*, il quale svolge un ruolo di vettore informativo per il sistema assumendo: valori uguali ad 1 nei casi in cui l'informazione risulta* x_i*, valori uguali a zero se l'evidenza espressa risulta essere connessa ad altri valori di x diversi da* X_i*, mentre nel caso in cui non vi fosse uno scambio*

di informazioni per X, questo sarebbe il prodotto di tutti i vettori informativi delta ricevuti dai relativi nodi figlio.
Il parametro $\pi(x_i)$, invece, rappresenta il prodotto della CPT e dei vettori π relativi ai nodi genitore.
Così facendo l'algoritmo permette di far collimare: le informazioni provenienti dai nodi figlio attraverso mediante il parametro $\lambda(X)$, i valori presenti della CPT e, infine, i π messaggi ricevuti dai nodi genitore.
L'algoritmo di passaggio informazioni si dimostra così in grado di esprimere le evidenze di ogni singolo nodo del sistema prescindendo dall'eventuale emersione di nuove evidenze.
Si precisa, inoltre, che la notazione $\pi - \lambda$ viene introdotta dai fautori dell'algoritmo e nello specifico si considera:

- *π il vettore informativo che segue la direzione degli archi, dai genitori ai figli.*

- *λ il vettore informativo che, contrariamente alle direzioni espresse dagli archi, si occupa dei passaggi informativi da figlio a genitore.*

2.3.2 Inferenza con informazioni incerte

Se fino ad ora abbiamo proceduto consapevoli dell'assunzione che ogni evidenza di nuova espressione nel sistema fosse il risultato di processo analitico in grado di formare un'informazione certa e precisa per la variabile esaminata, gli algoritmi di inferenza si dimostrano in grado di operare anche quando vi fosse un grado di incertezza connesso all'informazione, la quale verrà, quindi, definita come *virtuale* o di *verosimile*.
Sebbene questa non sarà la sede per una loro trattazione in dettaglio, varie tecniche vengono utilizzate per far fronte a questa problematica come per l'esempio: l'utilizzo di approcci inferenziali dotati di simulazioni stocastiche, semplificazioni logiche, pesature di verosimiglianza, la stipula di evidenze create virtualmente nonché il servirsi di algoritmi più complessi co0me la "Markov chain Monte Carlo.

2.3.3 Inferenza esatta in una rete a connessioni multiple

Quando ci si possa trovare in presenza di una situazione in cui due nodi sono connessi tra loro mediante più di un sentiero questa eventualità fa sì che la rete venga definita "a connessioni multiple" (si veda la figura 2.11 per una rappresentazione visiva).

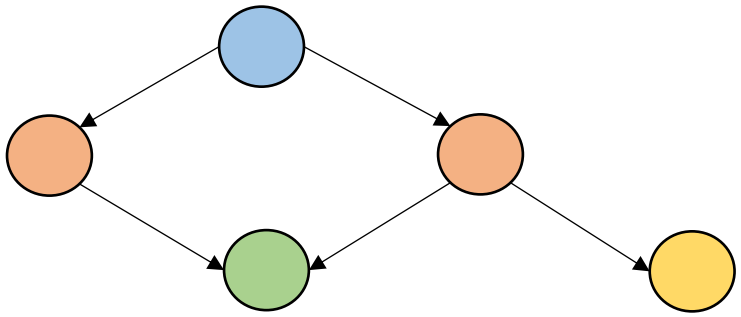


Figura 2.11: Diagramma di rete a connessioni multiple

In questa tipologia di network l'applicazione dell'algoritmo di passaggio informazione non trova applicazione poiché una variabile può essere influenzata da molteplici sentieri causali; per rispondere a questo problema si opera redigendo degli appositi algoritmi di *"clustering"*, i quali si dimostrano in grado trasformare la rete in un *"polytree"* mantenendo, però, inalterato il disegno probabilistico intercorrente tra le nodosità (figura 2.12).
Per rimuovere "i sentieri in eccesso" l'algoritmo si articola in due macrofasi:

1. Trasformazione del network in un polialbero, un passaggio forse computazionalmente lungo poiché in relazione alle dimensioni dei cluster da "razionalizzare" ma che tuttavia una volta compiuto non necessita di ulteriori revisioni.

2. Sviluppare il "*belief updating"* all'interno dell'appena creato polialbero.

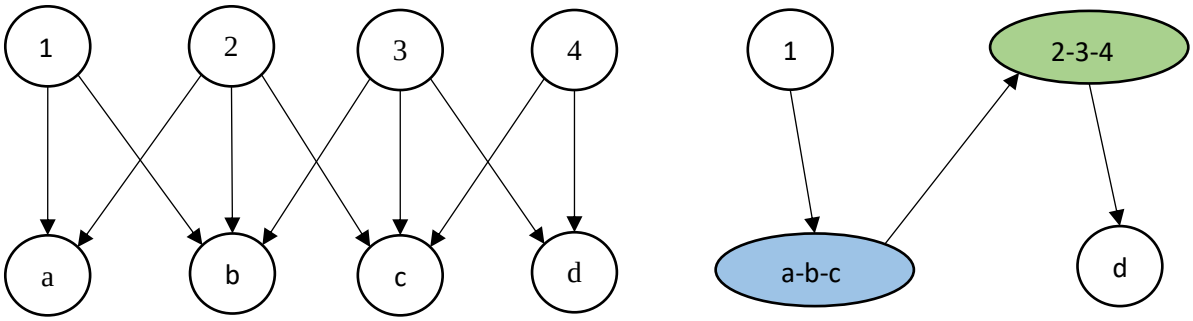


Figura 2.12: Esempio di "clustering" per la semplificazione dei sentieri causali e la creazione di un polialbero

Il processo inferenziale, a conclusione di questo processo di "merging", vede ora ritrovata la possibilità di applicare l'algoritmo di passaggio informazioni.

Alberi di giunzione

Il procedimento appena visto può essere nello specifico realizzabile attraverso l'applicazione del cosiddetto algoritmo di *"junction trees"*, il quale fornisce uno strumento efficiente per la semplificazione di una rete in un polialbero.

Algoritmo junction trees

È basato su una serie di passi di seguito riportati

1. *Moralize: questo passaggio prevede il congiungimento di tutti i nodi genitori, una sorta di matrimonio all'interno della struttura della rete, il risultato è anche detto "moral graph" (figura 2.13).*

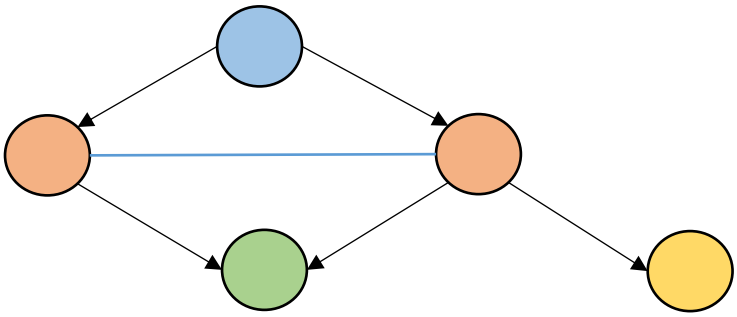


Figura 2.13: Rappresentazione del congiungimento dei nodi genitore

2. *Triangulate: vengono aggiunti archi qualora ogni ciclo di lunghezza superiore a tre abbia una corda al suo interno, questo può produrre diversi output di triangolazione in funzione del cluster preso in esame; il problema di triangolazione ottimale risulta quindi appartenente alla categoria dei problemi computazionali NP-completi e il grafo che ne si deduce è definibile "triangulated".*

3. *Create a new structure: questo step identifica il gruppo massimo di grafi triangolarizzati da unire in un unico nodo composito, i quali uniti poi tra loro formano il junction tree.*

4. *Create separators: ogni arco presente negli alberi di giunzione ha al suo interno un separatore che definisce le intersezioni dei nodi adiacenti.*

5. *Compute new parameters: ogni nodo e relativi separatori presenti nell'albero di giunzione hanno un'associata tavola di probabilità condizionata derivata da quella delle singole variabili in esse riassunte in modo che la CPT dell'albero di giunzione è uguale a quella della rete Bayesiana originale.*

6. *Belief updating: ora le nuove evidenze si possono propagare nell'albero di giunzione seguendo l'algoritmo di passaggio informazioni.*

2.4 Imparare la struttura delle reti Bayesiane: apprendimento della struttura

Lavorando con i network Bayesiani spesso si ricade nell'eventualità di non conoscere a priori la struttura della rete che codifica le distribuzioni di probabilità condizionata tra gruppi di nodi. È necessario dunque avere un insieme di dati riferiti al problema oggetto di studio per stimare la struttura della rete direttamente dai dati.
Per far fronte a questo problema son stati elaborati metodi di stima o "apprendimento" della struttura della rete, divisibili in due macrocategorie: algoritmi *search and score* e algoritmi *costraint-based*, i quali, sfruttando le potenzialità del *machine learning*,

ricavano la struttura partendo dalla base dei dati e dei relativi legami probabilistici a disposizione, adottando, comunque, gli opportuni accorgimenti logici.
Per quanto riguarda la metodologia "score based" essa si caratterizza per operare in maniera olistica; ovvero: partendo dai dati in possesso, sfruttando la potenza di calcolo a disposizione, questo tipo di approccio andrà a definire un insieme di strutture plausibili date le informazioni di partenza; successivamente, per ognuna delle possibili strutturazioni sarà associato un punteggio di merito, per valutarne le performance "topologiche"; infine, si aprirà una fase di confronto per esaminare quella tra le strutture individuate presenti il "fitting" migliore rispetto ai dati iniziali.
Gli approcci "costraint base", invece, seguono un diverso modo di affrontare il problema; a differenza di quelli "score", questa categoria si concentra sul valutare le relazioni di probabilità presenti tra le variabili di un dataset e, sulla base di queste, elaborare, poi, un network in grado di cogliere tutte le connessioni ottenute.
Gli approcci appartenenti alle due categorie non sono mutualmente esclusivi ciò, comporta ad un frequente uso combinato dei due, il che rende di conseguenza possibile identificarne una terza categoria, ovvero, quella: mista.
Nel seguito, verranno presentati gli approcci di apprendimento della struttura di rete Bayesiana più frequentemente utilizzati in letteratura.[9]

L'algoritmo K2 di Cooper & Herskovits

Tra gli approcci appartenenti alla categoria *search and score*, uno dei più rilevanti e significativi per affrontare questa tematica è stato presentato nel 1991 dai ricercatori Cooper ed Herskovits; il loro approccio, chiamato *K2*, propone di determinare la struttura di una rete, partendo da ipotesi individuali di probabilità condizionata, $P(h_i|e)$, per poi affrontare un problema di calcolo combinatorio.
In altri termini, l'obbiettivo di questo approccio è ricavare il parametro h_i in grado di massimizzare $P(h_i|e)$, il tutto partendo da un'iterazione del teorema di Bayes:

[9] Per una lettura più approfondita si faccia riferimento ai capitoli 8 dei libri: Kjaerulff & Madsen (2008), *Bayesian Network and Influence Diagrams*, Springer Science+Business Media, New York e Korb & Nicholson (2004), *Bayesian Artificial Intelligence*, Chapman & Hall/CRC, Boca Raton, Florida

$$P(h_i|e) = \frac{P(e|h_i)P(h_i)}{P(e)}$$
$$= \beta P(h_i, e)$$

Dove β rappresenta una costante di normalizzazione.

Per consentire al programma di operare correttamente è poi necessario fare delle assunzioni tali da ridurre la complessità di calcolo:

1. I dati sono "joint samples" e tutte le variabili sono discrete, così che:

$$\boldsymbol{P(h_i, e) = \int_{\theta} P(e|h_i, \theta) f(\theta|h_i) P(h_i) d\theta}$$

Dove: θ è il vettore parametro e $f(\cdot\,|h_i)$ è la densità a priori sui parametri della struttura casuale

2. I dati sono indipendentemente e identicamente distribuiti, ciò vale per k casi campione, mentre e_k è il riverbero dell'evidenza e all'interno dei suoi componenti K.

$$\boldsymbol{P(e|h_i, \theta) = \prod_{k=1}^{K} P(e_k|h_i, \theta)}$$

3. Il campionamento non presenta dati mancanti

4. Per ogni variabile X_k in h_i e per ogni istanza dei suoi nodi genitore $\pi(X_k)$, $P(X_k = x|h_i, \theta, \pi(X_k))$ è uniformemente distribuita sui possibili valori $X_k = x$

5. Si assume a priori per il modello spaziale la seguente relazione: $P(h_i) = \frac{1}{|\{h_i\}|}$

6. Si considera noto l'ordine temporale delle variabili

Sulla base di queste assunzioni è possibile, quindi, identificare la probabilità condizionata massimizzata attraverso la seguente teorema:

$$P_{CH}(h_i, e) = P(h_i) \prod_{k=1}^{N} \prod_{j=1}^{|\phi_k|} \frac{(s_k - 1)!}{(S_{kj} + s_k - 1)!} \prod_{l=1}^{s_k} \alpha_{kjl}!$$

Dove:

- N rappresenta il numero delle variabili
- $|\phi_k|$ indica il numero di possibili allocazioni di $\pi(X_k)$
- s_k indica il numero di possibili allocazioni di X_k
- α_{kjl} è il numero di casi nel modello per cui X_k prende l'l-esimo valore mentre $\pi(X_k)$ il j-esimo
- S_{kj}, infine, definisce il numero di casi nel modello tali che $\pi(X_k)$ prende il j-esimo

Con questo approccio, la definizione di una rete risulta, quindi, conseguenza di uno sforzo computazionale in grado di massimizzare il fitting globale del modello sulla base delle relazioni di probabilità condizionata.

BIC score approximation

Un ulteriore criterio per valutare lo score di una struttura di una rete partendo dai dati a disposizione è il cosiddetto metodo BIC; esso prevede, come per tutti gli algoritmi score-based, la massimizzazione del punteggio assegnabile al network candidato all'esplicitazione del dataset.

La relazione base per cui ciò possa risultare possibile prevede, quindi, di assegnare alla rete un punteggio in base ai dati D a disposizione e al modello di struttura testato, G:

$$Score(G, D) = P(G|D) = \frac{P(D|G)P(G)}{P(D)}$$

Appare relativamente chiaro che per massimizzare il punteggio in dipendenza alla struttura della rete risulta sufficiente lo studio del numeratore, per cui, apponendo l'assunzione di una distribuzione costate della probabilità a priori per $P(G)$, resta solo da analizzare: $P(D|G)$.
A questo punto del processo, per o studio di quest'ultimo fattore si procede con una media pesata di tutti i possibili parametri secondo la loro probabilità a priori:

$$P(D|G) = \int P(D|G, prior)P(prior|G)dprior$$

Questo, combinato con le evidenze sulla densità di probabilità ricavate dall'algoritmo K2 e mediante delle semplificazioni statistiche sulla distribuzione delle variabili e sulle relative medie, permette di ottenere il *BIC score* per la valutazione di un modello strutturale.

$$BICscore(G, D) = logP(D|\hat{p}, G) - \frac{d}{2}logN$$

Hill climbing

Un altro approccio di ricerca, combinabile con quelli appartenenti alla categoria score-based, per l'individuazione della struttura ottimale è chiamato "Hill climbing"; esso si basa sull'applicazione reiterata, detta anche "greedy search", di un algoritmo che, ad ogni iterazione porta all'individuazione del network migliore fino a trovare quello in grado di massimizzare il punteggio ottenibile dal calcolo di una possibile score.
In estrema sostanza, fatta salva la possibilità di utilizzare le conoscenze preliminari sul dataset per operare una preselezione dei network iniziali, l'algoritmo vede come operazioni principali: l'individuazione di una rete di partenza con il relativo punteggio,

poi, sulla base dello studio della densità di probabilità e delle varie relazioni di condizionamento, prevede l'apposizione o la rimozione singola di connessioni (archi nella rete), una volta fatto ciò, si individua il nuovo punteggio del network, e si compara con quello ottenuto in precedenza fino a che i punteggi non risultano più crescenti.[10]

Algoritmo PC

Appartenente alla categoria degli approcci *costraint-based,* questo algoritmo si presenta come una sorta di evoluzione dell'algoritmo IC di Verma e Pear, calmando alcune sue lacune riguardanti premesse poco realistiche.
Questo algoritmo si articola in sei fasi, le quali permettono di ricavare la struttura della rete a partire dalle informazioni dai rapporti di dipendenza e indipendenza tra le variabili.

Algoritmo:

1. *Iniziare con una rete "scheletro" in cui ogni nodo risulta adeguatamente collegato e connesso*

2. *Individuare un* $k \leftarrow 0$*; dove* k *identifica l'ordinamento del set di variabili da considerarsi fisse; successivamente, per tutte le coppie di nodi* X *e* Y *porre* $DSep(X,Y) = \emptyset$*: ciò terrà nota dei nodi, i quali andranno d-separati nel grafo finale.*

3. *Per ogni paio di nodi adiacenti* X *e* Y*, eliminare ogni arco tra loro solo ed esclusivamente se tutti i sottoinsiemi* S *di ordine* k *contenenti nodi adiacenti ad* X*, il modello di correlazione parziale non è significativamente diverso da zero; aggiungere, quindi, i nodi* S *al* $DSep(X,Y)$

4. *Se così facendo un arco venisse rimosso, si incrementi* k *e si ritorni al punto 3.*

[10] Per un approfondimento si legga il capitolo 4.5.1.2 del libro: Scutari & Denis (2015), *Bayesian Networks with examples in R,* Taylor & Francis Group, New York

5. *Per ogni gruppo di tre nodi X, Y e Z indirettamente congiunti ($X \longrightarrow Y \longrightarrow Z$), sostituire lo schema correlativo con il seguente $X \longrightarrow Y \longleftarrow Z$, se $Y \notin DSep(X, Y)$.*

6. *Applicare il terzo passaggio dell'algoritmo IC*

    a. *Per tutti gli archi indiretti $Y - Z$ presenti nel grafo porre tale orientamento $Y \longrightarrow Z$, solo ed esclusivamente se:*

        i. *Y appare come nodo centrale di una catena non direzionata con X e Z così che X e Y verrebbero direzionate in tal maniera: $X \longrightarrow Y$*

        ii. *Qualora vi fosse una connessione $Y \longleftarrow Z$ verrebbe introdotta una circolarità*

Per esigenze di sintesi, gli approcci qui presentati sono solo una ristretta cerchi di quelli a disposizione per la determinazione di un network Bayesiano; perciò si rimanda alla nota di pie pagina per dare una bibliografia di riferimento[11]

Dati incompleti

Gli approcci presentati per l'inferenza e l'apprendimento delle reti Bayesiane presuppongono il fatto che i dati su cui basare l'analisi siano completi. Spesso questa assunzione viene a mancare soprattutto quando si hanno a disposizione dati raccolti su fenomeni reali complessi.
Nello specifico in situazioni di dati mancanti la teoria Bayesiana permette di apportare alcune soluzioni efficaci affinché, partendo dai dati a disposizione, si possano colmare queste lacune e portare comunque a termine l'analisi.
Tra i principali metodi risolutivi proposti per questo problema si possono identificare due algoritmi: quello di Gibbs e l'approccio di massimizzazione delle aspettative.

[11] Si legga il capitolo 4.5 del libro: Scutari & Denis (2015), *Bayesian Networks with examples in R,* Taylor & Francis Group, New York

Nel seguito verranno introdotti sinteticamente i principi base che regolano questi due approcci nell'imputazione dei dati mancanti.

Gibbs sampling

Questo approccio è in grado di creare modelli sulla base di una qualsiasi funzione $f(X)$ ottenendo $P(f(X))$; in particolare, per affrontare il problema delle imputazioni mancanti, è possibile modellare la funzione di probabilità condizionata $P(X|e)$ dove l'evidenza e è solo parziale.
In termini più pratici, questo tipo di approccio permette di stimare P(X) da un qualsiasi punto iniziale, X, qualsiasi del network e di modellizzare, in seguito, un modello adiacente governato dalla funzione di probabilità condizionata P(X | e); l'utilizzo di tare apparato probabilistico permette di avere il controllo sulla distribuzione delle osservazioni cosicché, anche nel caso di network multinomiali, la distribuzione delle osservazioni può sopperire alle informazioni mancanti nei dati.[12]

Massimizzazione delle aspettative

Questo tipo di approccio, invece, segue una logica di tipo deterministico per stimare asintoticamente un network multinomiale, θ, data la presenza di osservazioni mancanti e l'assunzione che queste siano indipendenti rispetto a quelle effettivamente osservate.
L'algoritmo, quindi, restituisce una stima di $\hat{\theta}$ sulla base di θ ,la quale può alternativamente essere: o il risultato che massimizza la verosimiglianza di $P(e|\hat{\theta})$ o la massimizzazione della probabilità a posteriori di $P(\hat{\theta}|e)$.[13]

[12] Per approfondimenti si legga il capitolo 7.3.2.1 del libro: Korb & Nicholson (2004), *Bayesian Artificial Intelligence*, Chapman & Hall/CRC, Boca Raton, Florida

[13] Per una lettura più approfondita si suggerisce il capitolo 7.3.2.2 del libro: Korb & Nicholson (2004), *Bayesian Artificial Intelligence*, Chapman & Hall/CRC, Boca Raton, Florida

Algoritmo:

Premessa: Attribuire a un valore arbitrario "legale", indentificare un grado di precisione ϵ per $\hat{\theta}$; porre a $\hat{\theta}'$un valore eccessivamente elevato affinché esso sia un valore "illegale".

Mentre $\left|\hat{\theta} - \hat{\theta}\right| > \epsilon$ e $\hat{\theta} \leftarrow \hat{\theta}'$ (ad eccezione della prima iterazione) si calcoli:

1. *Expectation step: si determini la distribuzione di probabilità sui dati mancanti:*

$$P\left(e* \middle| e, \hat{\theta}\right) = \frac{P\left(e\middle|e*, \hat{\theta}\right)P(e* |\hat{\theta})}{\sum_{e*} P\left(e* \middle| e, \hat{\theta}\right) P(e* |\hat{\theta})}$$

2. *Maximization step: si computi $P(e|\hat{\theta})$ o alternativamente $P(\hat{\theta}|e)$ dati la stima di $\widehat{\theta'}$ e la relazione di probabilità $P\left(e* \middle| e, \hat{\theta}\right)$*

In conclusione, a prescindere dalla tecnica utilizzata, quando si affronta l'eventualità di imputazioni mancanti all'interno del dataset, il metodo migliore per ricavare le probabilità connesse alle lacune impositivi comporta il calcolo della distribuzione totale di probabilità definita dalla totalità dei parametri.

2.5 Ulteriori sviluppi delle reti Bayesiane

L'utilizzo delle reti Bayesiane non si limita ad operazioni di inferenza e rappresentazione probabilistica, bensì le loro proprietà possono essere mutuate per l'uso in campi applicativi molto eterogenei tra loro; uno di questi riguarda proprio la rappresentazione dell'utilità attesa derivate dalla creazione di modelli decisionali (anche detti diagrammi di influenza).

Alla base di questo campo applicativo si fanno saldi i principii economici di massimizzazione dell'utilità attesa e dell'agire razionale per l'elaborazione di una funzione di utilità comprensiva dell'apporto probabilistico-inferenziale.

Sulla base di queste considerazioni è stato possibile elaborare la seguente funzione di utilità attesa:

$$EU(A|E) = \sum_i P(O_i|E, A)U(O_i|A)$$

Dove:

- *E rappresenta le evidenze informative a disposizione dell'agente economico*
- *A identifica un'azione non deterministica con il possibile stato risultante* O_i
- $U(O_i|A)$ *l'utilità di ogni possibile stato data pe compiuta l'azione A*
- $P(O_i|E, A)$ *la distribuzione di probabilità condizionata sul possibile stato risultante data un'evidenza* E *e un'azione A.*

Reti decisionali

Questa tipologia di network di derivazione Bayesiana, come suggerisce il nome, è in grado di rappresentare e porre considerazioni in merito al processo decisionale con una valutazione congiunta di alcuni input quali: lo status quo, l'azione e le decisioni da porre a confronto con i rispettivi esiti in termini di utilità attesa.
Per rendere più chiare e funzionali questa tipologia di rete si aggiunge un nuovo lessico topografico che pone una differenza rappresentativa tra i nodi della rete in funzione della tipologia di informazione che essi esprimono (figura 2.14):[14]

[14] Per degli esempi in letteratura si faccia riferimento al capitolo 4 del libro: Korb & Nicholson (2004), *Bayesian Artificial Intelligence*, Chapman & Hall/CRC, Boca Raton, Florida

Figura 2.14: Differenze rappresentative dei nodi appartenenti a reti decisionali

- Nodi causali (*chance nodes*): di forma ovale, rappresentano le variabili casuali a cui è possibile associare le relative *CPT* come per le più classiche reti Bayesiane, i loro nodi genitori possono essere nodi decisionali o ulteriori nodi causali.

- Nodi decisionali (*decision nodes*): di forma rettangolare, rappresentano i valori alternativi delle possibili decisioni prese dall'agente in un particolare istante temporale, un network decisionale con presente al suo interno un'unica decisione avrà di conseguenza un unico nodo decisionale, qualora nel network ci siano più decisioni in sequenza i nodi decisionali possono avere come nodi genitori ulteriori nodi decisionali, rispecchiando la natura di un processo decisionale sequenziale; qualora vi fosse un nodo causale come genitore di un nodo decisionale è da considerarsi data l'informazione all0interno del nodo causale nell'istante in cui si compie la decisione.

- Nodi utilità (*utility nodes o value nodes*): di forma a diamante, questa tipologia di nodi rappresentano la funzione di utilità attesa dell'agente economico dando, perciò, informazioni sulle utilità attese ai diversi sentieri decisionali; ad ogni nodo probabilità è associata una "tavola di utilità" che riassume tutti gli input provenienti dai nodi genitori, in presenza di molteplici nodi utilità l'utilità totale è semplicemente la sommatoria delle singole utilità.

Valutazione di un network decisionale

È possibile valutare un network decisionale in cui è presente un unico decision node applicando il seguente algoritmo:

1. *Aggiungere ogni informazione a disposizione*
2. *Per ogni valore nel nodo decisionale:*
    - *Stabilisci quel valore come valore del nodo*
    - *Calcola le probabilità a posteriori dei nodi genitore del nodo utilità utilizzando un algoritmo inferenziale standard*
    - *Calcola la possibile utilità attesa dell'azione*
3. *Restituisci l'azione con la maggior utilità attesa.*

Quando ci si trova di fronte a strutture più articolate e complesse per il processo di valutazione risulta necessario far ricorso a una rappresentazione in grado di elaborare la maggior mole di informazione, per far ciò si vede, quindi, l'utilizzo di alberi decisionali (decision tree, Figura 2.15).
In questa modellazione i nodi intermedi e radice possono essere sia nodi decisionali sia nodi causali mentre i nodi foglia rappresentano esclusivamente le utilità attese del sistema; il linguaggio topografico si mantiene inalterato da quello utilizzato per i più semplici network decisionali; per ogni decision node è poi aggiunta un'indicazione sulle possibili alternative mentre per ogni nodo causale l'indicazione aggiunta riguarda tutti i possibili valori esprimibili dal nodo.
Ulteriore aggiunta portata da questo tipo di modellazione è il principio di non dimenticanza, il quale stabilisce che l'agente del processo decisionale interrogato su un qualsiasi punto del sentiero all'interno del network, esso ricorda tutte le indicazioni precedenti che lo hanno portato a quel punto.

Inoltre, ad ogni arco scaturito da un nodo causale è associata la relativa probabilità mentre ad ogni nodo foglia è associata la relativa utilità attesa emersa dai valori di tutte le indicazioni presenti nel sentiero che ha condotto a lei.

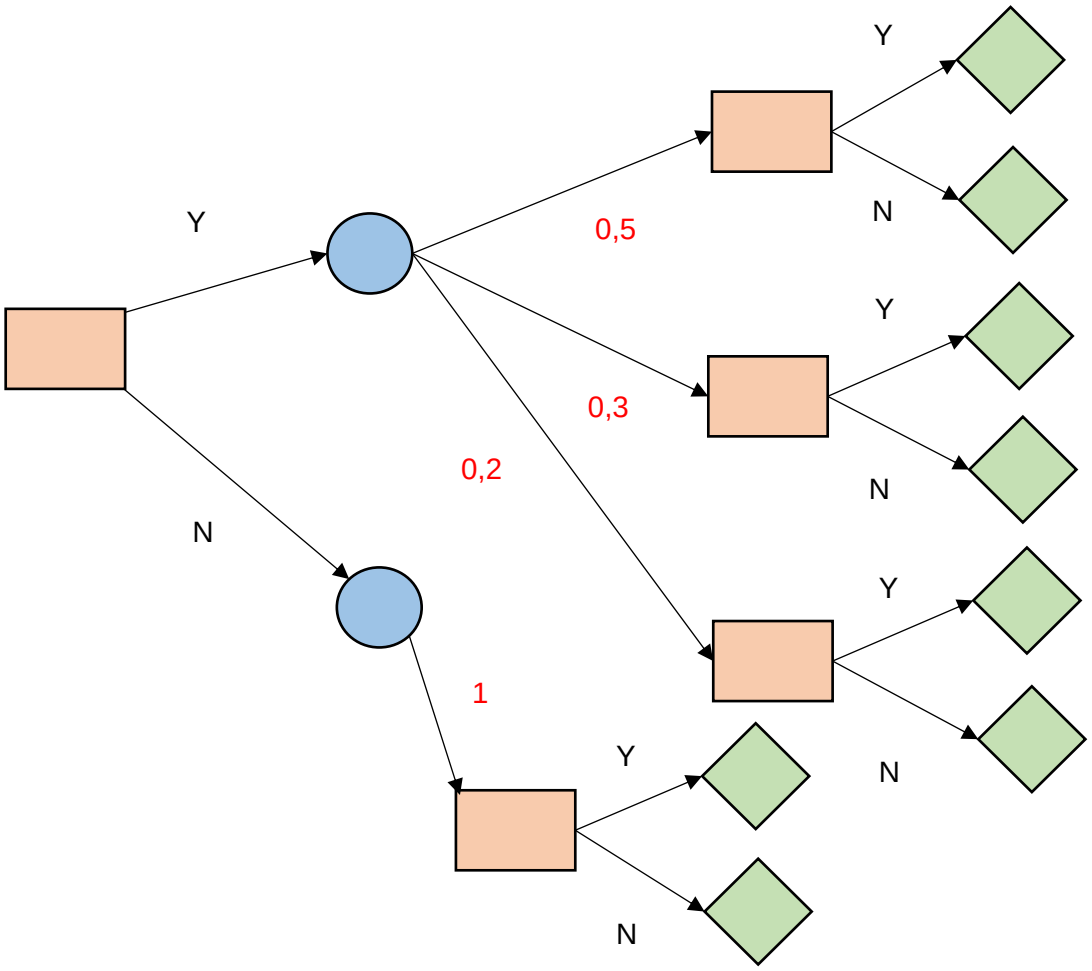


Figura 2.15: Rappresentazione di un albero decisionale, nel quale ad ogni nodo causale è associato un arco indicante probabilità di ottenimento del successivo nodo decisionale

Come per il caso in cui nel network si riscontri un unico nodo decisionale anche il processo di valutazione di un albero decisionale risulta attuabile mediante un algoritmo che prevede:

1. *Iniziare dai nodi aventi come figli solo nodi foglia*

2. *Qualora il nodo X fosse un nodo causale ogni arco in uscita ha espressa la sua relativa probabilità e ogni nodo figlio la sua utilità attesa. Per il calcolo dell'utilità attesa si utilizza la seguente formula:*

$$EU(X) = \sum_{C \in Children(X)} U(C) \times P(C)$$

In seguito, verrà poi preferita la decisione che massimizza l'utilità attesa

$$EU(X) = max_{C \in Children(X)}(EU(C))$$

3. *Ripetere poi ricorsivamente il processo ad ogni livello del network, utilizzando la funzione di utilità attesa per ogni nodo figlio*

4. *Il valore del nodo radice è la massima utilità attesa se ad ogni livello decisionale si massimizza l'utilità ottenibile.*

Reti Bayesiane dinamiche

Sebbene le reti decisionali e quelle Bayesiane rappresentino "uno stato del mondo" ad un particolare istante dell'arco temporale sotteso ad una determinata decisione o fenomeno di analisi; questa tipologia di network non esplicita in modo chiaro le relazioni temporalmente dinamiche che possono intercorrere tra le variabili.[15]
L'unica soluzione possibile a questo tipo di problema prevede l'inserimento di ulteriori variabili nella rete affinché esse interpretino le relazioni tra il dominio corrente delle nodosità con le rispettive versioni passate o future.
Si suppone quindi di avere un modello composto da n variabili casuali $X = [X_1, \dots, X_n]$ ognuna di queste associabile ad un nodo nel network Bayesiano in riferimento, la

[15] Per maggiori informazioni si legga il capitolo 4.5 del libro: Korb & Nicholson (2004), *Bayesian Artificial Intelligence*, Chapman & Hall/CRC, Boca Raton, Florida

componente dinamica per la rete si instaura includendo per ogni singolo nodo le versioni temporalmente posteriori X_{t+1} e temporalmente anteriori X_{t-1}:

I. Corrente: $\{X_1^t, X_2^t, \dots, X_n^t\}$
II. Anteriore: $\{X_1^{t-1}, X_2^{t-1}, \dots, X_n^{t-1}\}$
III. Posteriore: $\{X_1^{t+1}, X_2^{t+1}, \dots, X_n^{t+1}\}$

Ogni "arco temporale" prende il nome di *"time-slice"* e le relazioni tra le variabili appartenenti al medesimo arco si definiscono *"intra-slice arcs"* ($X_1^t \longrightarrow X_2^t$), mentre quelle che intervengono a istanti di tempo diversi *"inter-slice arcs"* ($X_1^{t-1} \longrightarrow X_1^{t+1}$); in molti casi le evidenze espresse per il livello temporale si propagano nella variabile anche agli istanti successivi, tuttavia, si mantiene generalmente salda la struttura del network all'interno del medesimo time-slice.

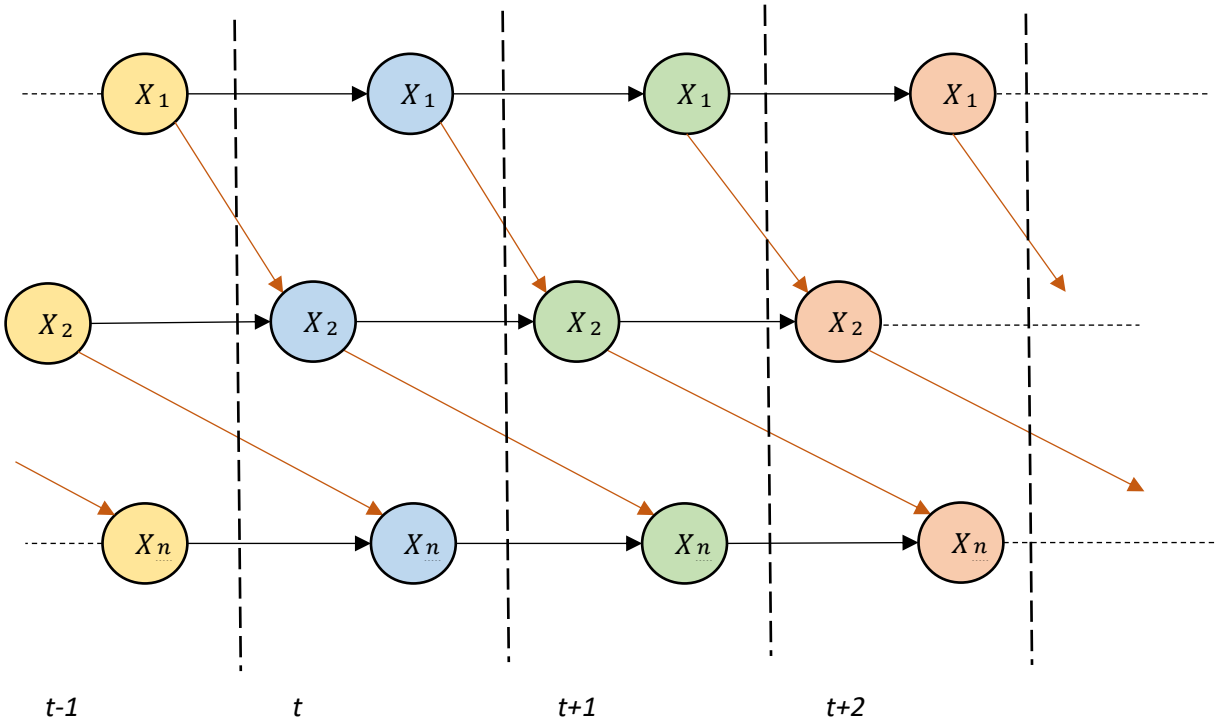


Figura 2.16: Rappresentazione di una rete dinamica con in evidenza gli archi inter e intra slice

La figura 2.16 esplicita a livello grafico il fluire di legami all'interno di una rete dinamica permettendo di far notare come all'interno del medesimo arco temporale non vi siano

archi che saltano più di una fascia temporale; questo risulta essere un ulteriore esempio connesso alle assunzioni di Markov, per cui un determinato istante temporale dipende solo dai precedenti stati e dalle azioni in essi compiute.

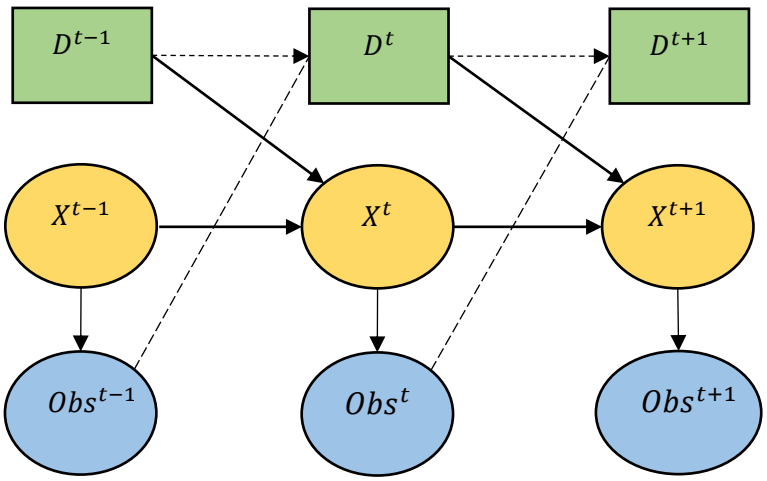
A livello formale le relazioni presenti nella *CPT* tra variabili inter e intra slice si desumono attraverso l'applicazione della seguente equazione di probabilità condizionata:

$$P(X_i^T | Y_1^T, \dots, Y_m^T, X_i^{T-1}, Z_1^{T-1}, \dots, Z_r^{T-1})$$

Dove: X_i^T *è un nodo generico con* $Y_1^T, \dots, Y_m^T$ *genitori intra-slice e* X_i^{T-1} *e* $Z_1^{T-1}, \dots, Z_r^{T-1}$ *genitori inter slice.*

Reti decisionali dinamiche

Analogamente con quanto fatto per i network Bayesiani anche le reti decisionali possono essere assoggettate al fluire delle dinamiche temporali (figura 2.18), cosicché si possono esplicitare non solo i mutamenti del dominio dovuti al passare del tempo ma anche gli effetti dovuti a processi di scelta sequenziali.[16]



[16] Per approfondire si faccia riferimento al capito 4.6 del libro: Korb & Nicholson (2004), *Bayesian Artificial Intelligence*, Chapman & Hall/CRC, Boca Raton, Florida

Figura 2.18: Esempio di rete decisionale dinamica

La figura in questione descrive una generica rete di decisione dinamica per eseguire n decisioni $D^t, \dots, D^n$; l'iter decisionale e il fluire del tempo sono individuati attraverso l'insieme di legami definito dagl'archi, ogni singolo nodo causale *X* individua una precisa utilità, la quale attraverso le evidenze del nodo osservazione (*Obs*) la rende informazione utile per la successiva decisione; questo processo si ripete per tutto il fluire temporale sotteso al processo decisionale rispettando il vincolo di massimizzazione dell'utilità per tutti gli n passaggi.

Reti Bayesiane Object Oriented

Modelli complessi caratterizzati da grandi moli di dati spesso presentano al loro interno sub-strutture ricorsive o medesimi componenti all'interno del dominio di riferimento. Data la premessa, trovano, in questo frangente, applicazione i network *object oriented*, o più semplicemente *Obj*, in quanto meccanismi semplificanti il modello d'analisi in grado di far emergere strutturazioni gerarchiche e di supportare la costruzione del modello.[17] Nei network probabilistici *Obj* vi è la presenza di "oggetti" costruiti sulla base di determinate variabili e archi relazionali, questo implica che nella diramazione topografica della struttura, a fianco delle classiche nodosità, il grafo *Obj* mette in luce legami tra nodi presenti in luoghi diversi all'interno del medesimo modello in analisi (figura 2.19).

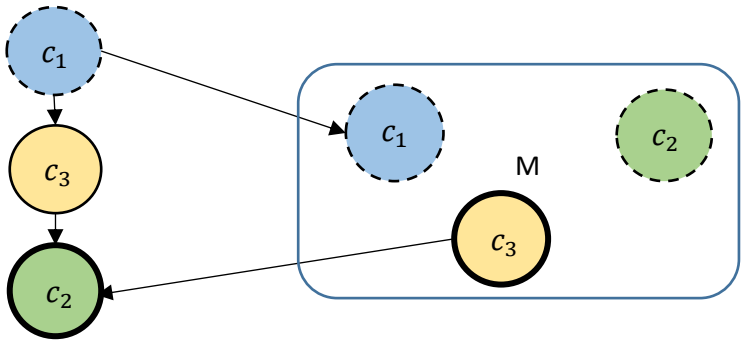


Figura 2.19: Esempio di rete *Obj*

[17] Per una lettura più approfondita si faccia riferimento al capitolo 4.3 del libro: Kjaerulff & Madsen (2008), *Bayesian Network and Influence Diagrams*, Springer Science+Business Media, New York

Nella rappresentazione grafica di una rete *Obj*, i legami causali emergenti, che altrimenti sarebbero rimasti celati all'interno del network, vengono raccolti e delimitati da sezioni di piano rettangolari dagli angoli smussati, al loro interno le variabili causali o di input vengono identificate da nodi tratteggiati mentre i nodi evidenza o output si distinguono per i contorni in grassetto.
Più formalmente una rete *Obj* può, infine, essere definita come un network di classe $C = (N, I, O)$ in cui:

- N rappresenta una rete probabilistica su X variabili mediante un strutturazione *DAG* G

- I individua un gruppo di variabili base $I \subseteq X$ specificate come variabili input, mentre $O \subseteq X$ identifica un set di variabili output tali che $I \cap O = \emptyset$ e $H = {}^{X}/_{(I \cup O)}$

2.6 Considerazioni finali su modello di rete Bayesiana

In conclusione, dopo questa breve introduzione sulla teoria dei grafi e la presentazione delle principali proprietà dei network Bayesiani, appare chiaro come questi si siano rivelati strumenti analitici molto efficaci e duttili.

La teoria della probabilità e l'approccio Bayesiano risultano, inoltre, utili per confrontarsi con lo studio dell'incertezza in approcci tipici del *machine learning*.

La facilità con cui sono in grado di affrontare la computazione di complesse relazioni di probabilità condizionata all'interno dei dataset, la capacità di una rappresentazione visiva e concreta dei legami intercorrenti tra le variabili, il novero di algoritmi a disposizione per fronteggiare le varie necessità analitiche, da quelle di natura inferenziale fino a quelle di natura topologica, l'essere in grado di sfruttare il machine learning per l'autodeterminazione della proprio struttura, prescindendo anche dalla presenza di informazioni mancanti, e ,in ultimo, la grande abilità di comprendere, e rappresentare, a pieno le dipendenze e le indipendenze delle variabili rilevate nei dati, imprimendone

anche una direzione a questi, fanno sì che l'utilizzo di queste reti sia in grado di supportare a pieno un analisi statistica come quella che si sta cercando di sviluppare.

Capitolo 3
Le reti Bayesiane come strumento di rappresentazione e analisi in ambito turistico

Nel presente capitolo verranno introdotti i dati usati come caso studio per descrivere le peculiarità dello strumento di rete Bayesiana, mostrando come i risultati ottenuti possano essere d'aiuto nello spiegare le relazioni presenti tra le variabili rilevati per un particolare fenomeno d'interesse e fornire supporto alle decisioni.

3.1 I dati

Il dataset considerato è parte dei microdati disponibili a libero uso nel sito dell'Istat[18]. Essi riguardano, nello specifico, un'indagine effettuata dall'Istat per scoprire il comportamento degli italiani nelle loro scelte turistiche, dove per turismo si intende: "l'insieme delle attività e dei servizi riguardanti le persone che si spostano al di fuori del loro 'ambiente abituale', per vacanza, motivi lavorativi o anche di salute[19]; sebbene la definizione si estende per una molteplicità di motivi, nel contesto di analisi ci si limiterà a tener conto soltanto delle rilevazioni pertinenti all'ambito puramente vacanziero, tralasciando, quindi, tutte le osservazioni connesse a spostamenti per fini lavorativi o di salute.

Su queste basi si andranno, perciò, a prendere e confrontare i dati emersi dai questionari del 2020, 2019, 2018 e 2014 con lo scopo di esaminare come di anno in anno emergano relazioni tra le variabili presenti, sia in un'ottica di breve periodo che di medio termine grazie alla presenza delle rilevazioni dell'anno 2014, le quali permettono di valutare più compiutamente come si è evoluto questo fenomeno nel tempo.

L'analisi descrittiva intertemporale permetterà, quindi, di poter ottenere maggiori informazioni sui key-drivers del turismo e di valutare come questi possono essere, eventualmente, mutati nel tempo; la rapidità e la crescente diffusione di nuovi approcci mediati dall'incalzante sviluppo tecnologico rendono settori come questo molto fluidi nel tempo, l'impatto di servizi come AirB&B, con la sua crescente popolarità, e di possibili

[18] Si apporta di seguito l'URL che indirizzerà alla pagina dell'Istat in cui sono messi a disposizione le campionature qui analizzate e le informazioni a loro corredo: https://www.istat.it/it/archivio/178695
[19] Si mette a disposizione l'URL che conduce agli aspetti metodologici utilizzati dall'Istat: https://www.istat.it/microdata/download.php?id=/import/fs/pub/wwwarmida/264/2020/01/Nota.pdf

future innovazioni contribuiscono a far diventare questo settore tanto stabile nel tempo, vedasi la forte stagionalità dei flussi turistici, quanto mutabile negli approcci, la facilità di molte applicazioni di booking ha reso talvolta obsolete figure quali agenzie o tour operator.

I dati si riferiscono a informazioni su un campione di famiglie, di fatto e non di diritto, residenti sul suolo italiano e dai soggetti che le compongono.

Il campionamento avviene tramite delle liste di selezione ricavate dalle anagrafiche comunali, anche dette LAC, successivamente, da ogni lista vengono poi estratti dei sotto partizionamenti di famiglie che definiranno le unità finali di campionamento.

Le informazioni riguardanti le abitudini vacanziere vengono poi ricavate attraverso un apposito focus presente nell'intervista dell'indagine sulle Spese delle Famiglie.

L'intervista è di tipo diretto e viene condotta con l'approccio Capi (intervista faccia a faccia assistita da computer); ogni componente della famiglia viene, perciò, intervistato sulle vacanze effettuate durante il periodo d'indagine il quale si estende per tutta la durata dell'anno così da cogliere le stagionalità del turismo.

In tutti gli anni portati in analisi il modello di campionamento del dataset si mantiene costante ad esclusione dell'anno 2014 in cui vengono a mancare alcune caratteristiche invece rilevate per gli anni successivi. I dati analizzati si compongono quindi per gli anni dal 2020 al 2018 di 52 variabili e di 46 variabili per il 2014, osservate su un campione di 3366 osservazioni per il 2020, 4393 per il 2019, 4705 per il 2018 e 3413 per il 2014.

3.2 Pre-processing dei dati

"A data set is a collection of data that describes attribute values (variables) of a number of real-world objects (units)".

In situazioni come questa, quando ci si trova a lavorare con dati reali, prima di avviare il processo di analisi sui dati raccolti è spesso necessario una fase preliminare di pulizia del dato, anche in termini di selezione od organizzazione delle variabili più rilevanti per il sistema oggetto di studio.

Congiuntamente alla selezione delle variabili, occorre porre le dovute attenzioni anche sulla natura qualitativa e quantitativa dei dati raccolti; questi, per, l'appunto, si presentano inizialmente come dei semplici *raw* data (figura 3.1) perciò, come per quanto fatto per le variabili, ci si aspetta di operare anche per i dati una serie di processi di "pulizia" di modo che siano pronti per le successive applicazioni analitico-statistiche.

Per definizione i *raw* data esibiscono notevoli lacune nell'inquadramento dell'informazione raccolta sul framework di campionamento statistico; dati mancanti, imputazioni non corrette o categorizzazioni imprecise rendono di conseguenza il dataset inadatto per l'applicazione di strumenti statistici quali le reti Bayesiane.

Affinché si possa avere un base informativa solida per la futura analisi risulta essenziale la trasformazione dei dati da *Raw* a *Consistent*; per fare ciò è necessario operare

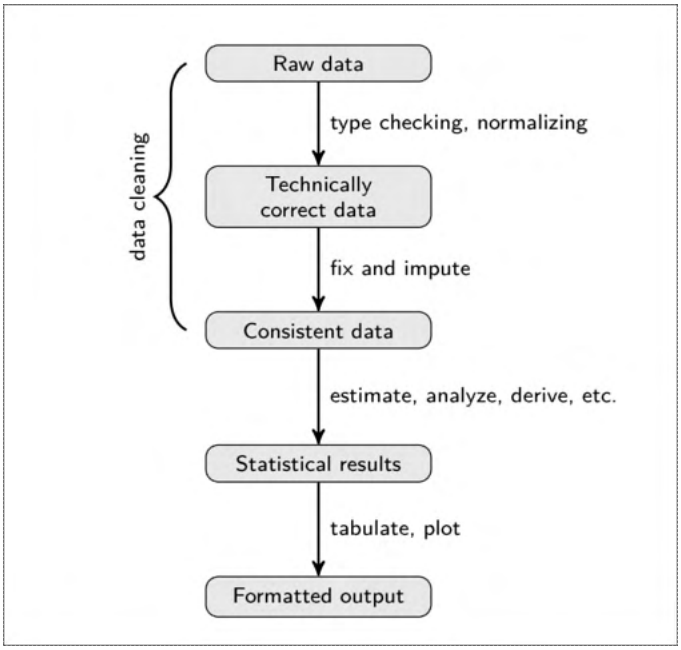


Figura 3.1: Diagramma di processo per la pulizia dei dati

opportune azioni di pulizia affinché poi l'output rispecchi veramente la conoscenza concreta desunta dalla campionatura, prescindendo dalla presenza di errori al suo interno.

Nello specifico, in relazione alle imputazioni mancanti, forse la forma più evidente di errore all'interno della base dati, la disciplina suggerisce diversi iter per la corretta gestione di questi:

I. La diretta eliminazione di questi

II. *Il recupero dell'informazione effettuando un parziale indagine sulle fonti della campionatura*
III. *L'imputazione diretta di questi: attraverso l'esplicitazione di alcune relazioni presenti tra le variabili (ad esempio una rilevazione mancate alla voce salario mensile può essere ricavata moltiplicando il salario orario per il numero di ore lavorate nel mese), copiando informazioni presenti in rilevazioni simili, mediante l'utilizzo di processi di stima*

Al termine della pulizia sul campione *raw*, il dataset è pronto per un'analisi di tipo statistico: le osservazioni mancanti son state rimosse o ripristinate, gli errori son stati corretti o eliminati e gli *outliers* giudicati rilevanti o meno, in quanto la loro presenza non è da considerarsi sbagliata, ma, bensì, deve essere valutata secondo criteri di pertinenza delle informazioni da loro espresse con i fini dell'analisi.
Inoltre le informazioni presenti dovranno essere prive di contraddittorio (un esempio di ovvia inconsistenza è la registrazione per un soggetto di età inferiore agli 8 anni di un posto lavorativo o di un titolo di studio) affinché l'informazione finale si possa anche definire pienamente "*consistent*".
Per fare ciò si applicano tre ulteriori passaggi al dataset, che prevedono: l'individuazione dell'incoerenza, la determinazione delle sue cause e, infine, la correzione formale delle imputazioni in tal modo individuate.
Così facendo, al termine di tutti questi passaggi, come rappresentato anche nella Figura 3.1 si completano le operazioni sulla base dati, in tal modo, le rilevazioni, così selezionate, rendono possibile un'analisi statistica il cui output è pienamente giustificabile sia in termini di correttezza analitica che di piena coerenza con le informazioni presenti nel dataset.
Per il caso studio preso in esame, si tiene a precisare che tutte le azioni svolte come pre-processing dei dati sono state applicate in egual maniera agli anni di campionamento considerato: il rispetto di tale coerenza logica risulterà propedeutico per un pieno confronto intertemporale, in quanto non vi è discriminazione nel trattamento dei vari dataset.
Nello specifico, la trasformazione dei dati da *raw* a *consistent* prevede, innanzitutto, l'eliminazione di una serie di variabili che forniscono informazioni superflue o trascurabili ai fini dell'analisi.

Sono stati esclusi tutte le osservazioni riguardanti i soggetti intervistati di età inferiore ai 14 anni; questo approccio ha un doppio beneficio per la base dati comportando sia l'eliminazione di numerosi NA sia la restituzione di un campione formato da persone con reale indipendenza economica e autonomia decisionale in ambito vacanziero.
Successivamente, si sono tenuti in considerazione tutte le osservazioni rilevanti viaggi con fini vacanzieri eliminando i viaggi connessi a scopi lavorativi.
Inoltre, dal dataset è stato possibile identificare se il soggetto intervistato ha sostenuto più viaggi durante l'anno di analisi, variabile che è stata creata e aggiunta per aumentare l'informazione sui comportamenti di viaggio degli intervistati.
Al termine di questi processi di pulizia si è ottenuto, per i quattro anni in esame, una numerosità campionaria notevolmente ridotta rispetto a quella di partenza, tuttavia però, l'informazione presente al suo interno, mancando di incoerenze tanto oggettive quanto logiche nell'imputazione dei dati e vista la totale assenza di NA, è, quindi, da considerarsi "consistent" e pronta per i successivi step di analisi descrittiva.
Nello specifico, si avranno 2497 osservazioni per l'anno 2014, 2371 per il 2018, 2292 per il 2019 e 1896 per il 2020.

3.3 Analisi descrittiva

Una volta terminato il processo di pulizia dei dati campionari e avendo perciò ottenuto una base di dati consistente e coerente si procederà ora applicando una disamina di statistica descrittiva sulle rilevazioni presenti nelle variabili oggetto di studio, facendo attenzione anche ad eseguire il tutto in una logica di confronto intertemporale.
Per le analisi, qualora un soggetto abbia effettuato più vacanze ci saranno molteplici interviste a suo nome, tuttavia in questa tesi date rilevazioni si assumeranno come non correlate; quindi, come se le vacanze e le relative interviste a riguardo fossero eseguite da soggetti diversi; questa scelta è stata fatta in previsione dell'elaborazione con le reti Bayesiana dei dati raccolti in quanto assunzione necessaria per la sua applicazione.

a.	Percentuale di soggetti che ha effettuato più viaggi nel corso dell'anno							
	2014		2018		2019		2020	
SI	7,31%	↘	6,21%	↘	3,82%	↗	5,31%	
NO	92,69%	↗	93,79%	↗	96,18%	↘	94,69%	

b.	Totale dei soggetti intervistati							
	2014		2018		2019		2020	
	2299	↘	2222	↘	2200	↘	1790	

Tabelle a. e b.: Le tabelle a. e b. restituiscono le informazioni sulle percentuali di intervistati che hanno effettuato più viaggi nel corso dell'anno di indagine e sul numero effettivo di soggetti intervistati.

Le tabelle a. e b. specificano quanto appena detto facendo intendere come il numero degli intervistati effettivi sia diverso dal quello ottenibile, per esempio, dalla somma delle categorizzazioni in fasce d'età; il numero degli effettivi soggetti intervistati risulta ottenibile solamente pulendo il campione dalle plurime rilevazioni effettuate in capo a chi ha compiuto più vacanze, che, come si vede nella tabella a., risulta essere comunque una minoranza del campione in trend apparentemente decrescente se si esclude il 2020.

Descrizione anagrafica

1.	Sesso							
	2014		2018		2019		2020	
Masch.	1173	46,98%	1123	47,36%	1084	47,29%	890	46,94%
Femm.	1324	53,02%	1248	52,64%	1208	52,71%	1006	53,06%

Tabella 1: La tabella presenta le statistiche assolute e relative sul sesso dei soggetti intervistati.

La distribuzione sul sesso degli intervistati si mantiene in costante decrescita in termini numerici di soggetti raggiunti, questo poi si rifletterà anche su ogni variabile che andremo ad esaminare. In ogni caso, per tutta la durata dei rilevamenti si mantengono, invece; perfettamente stabili gli equilibri percentuali tra le due controparti.

2.	Età							
	2014		2018		2019		2020	
15-24	289	11,57%	301	12,70%	297	12,96%	215	11,34%
25-34	303	12,13%	267	11,26%	287	12,52%	218	11,50%
35-44	521	20,87%	481	20,29%	356	15,53%	376	19,83%
44-54	510	20,42%	515	21,72%	504	21,99%	428	22,57%
55-64	456	18,26%	428	18,05%	443	19,33%	348	18,35%
65-74	307	12,29%	278	11,73%	277	12,09%	237	12,50%
over 75	111	4,45%	101	4,26%	128	5,58%	74	3,90%

Tabella 2: La tabella presenta le statistiche assolute e relative all'età dei soggetti intervistati, la quale viene suddivisa in sette scaglionamenti.

Per quanto riguarda le rilevazioni sull'età ci si trova di fronte a piccole variazioni nei pesi delle varie fasce generazionali; tuttavia, la portata di queste è da considerarsi irrilevante; risulta apprezzabile il fatto che il fulcro degli intervistati è principalmente orientato nella macro-fascia 35-64 con un assottigliamento percentuale, seppur lieve, della coda relativa alle popolazioni più giovani.

3.	Stato civile							
	2014		2018		2019		2020	
celibe/nubile	780	31,24%	794	33,49%	807	35,21%	649	34,23%
coniugato	1404	56,23%	1326	55,93%	1222	53,32%	1022	53,90%
separato/divorziato	200	8,01%	175	7,38%	182	7,94%	169	8,91%
vedevo	113	4,53%	76	3,21%	81	3,53%	56	2,95%

Tabella 3: La tabella restituisce le evidenze ottenute sullo stato civile dei soggetti intervistati, sia in valore assoluto che relativo.

La distribuzione relativa allo stato civile degli intervistati si mantiene anch'essa stabile nei partizionamenti, le variazioni seppur presenti non rendono apprezzabile alcun trend o modificazione delle variabili in oggetto (si fa salva anche qui l'omogeneità della base campionaria).

4.	Paese di nascita							
	2014		2018		2019		2020	
Italia	2318	92,83%	2203	92,91%	2153	93,94%	1776	93,67%
estero	179	7,17%	168	7,09%	139	6,06%	120	6,33%

La tabella 4: La tabella presenta le statistiche dall'analisi del paese di nascita dei soggetti intervistati, riportando sia le misurazioni assolute che relative.

La medesima considerazione la si apporta anche nella valutazione del paese di nascita che per tutti gli anni rilevati non prescinde dal conservare una percentuale di soggetti di provenienza straniera ad arricchire il risultato prodotto dal censimento.
Conclusa questa breve disamina limitata al contesto anagrafico d'ora in avanti si procederà per tutte le variabili di successiva analisi prescindendo dal fornire una traccia quantitativa sul numero di soggetti rispondenti una data categoria, come su fatto, ma ci si limiterà a prestare un'indicazione puramente percentuale in quanto si è interessati alle possibili variazioni pesate dei vari records.

Classificazione Geo-Culturale

5.	Ripartizione geografica di residenza						
	2014		2018		2019		2020
Nord-Ovest	25,03%	↗	28,76%	↗	29,80%	↘	29,27%
Nord-Est	32,68%	↘	30,20%	↗	33,94%	↗	34,02%
Centro	20,22%	↗	24,50%	↘	21,99%	↗	24,53%
Sud	18,74%	↘	11,51%	↘	10,03%	↘	8,76%
Isole	3,32%	↗	5,02%	↘	4,23%	↘	3,43%

La tabella 5: La tabella restituisce la distribuzione relativa delle residenze dei soggetti intervistati.

A differenza del paragrafo precedente da qui l'analisi porrà maggior attenzione all'andamento dinamico delle variabili poste in esame e ciò lo si può subito riscontrare grazie alla tabella relativa al partizionamento geografico degli intervistati, il quale fa emergere un mutamento intercorso nei vari anni di rilevazione: risulta visibile, in vero, un apprezzabile diminuzione di soggetti intervistati provenienti dal sud Italia, meno evidente nei confronti delle isole, in favore di un concentramento campionario tra centro e nord Italia.

La natura di questo fenomeno al momento non risulta chiara, le cause possono essere molteplici, ci si può limitare all'ipotesi di una possibile perdita di rappresentatività statistica per i comuni del sud dovuta forse all'emigrare dei cittadini verso le città più grandi o il nord Italia.

6.	Istruzione						
	2014		2018		2019		2020
nulla/diploma elemer	6,29%	↘	3,29%	↘	2,44%	↘	2,06%
licenza media/avv. Pr	21,63%	↘	19,78%	↗	21,07%	↘	19,62%
dpl. Super. o qual pro	44,29%	↗	46,31%	↘	45,94%	↗	46,41%
dpl univ. e Post laurea	27,79%	↗	30,62%	↘	30,54%	↗	31,91%

La tabella 6: La tabella mostra i vari livelli di istruzione rilevati dai dataset a disposizione fornendo un'indicazione dell'andamento tra i vari anni rappresentati.

Dal punto di vista dell'istruzione appare una visibile riduzione delle fasce "più deboli" in favore di un modesto aumento percentuale di intervistati possedenti un diploma superiore equiparabile qualifica professionale nonché titoli di studio quali la laurea; segno di un lodevole arricchimento culturale della base campionaria.

7.	Condizione professionale soggettiva						
	2014		2018		2019		2020
occupato	53,66%	↗	60,78%	↘	58,60%	↗	61,02%
in cerca lav.	7,13%	↘	4,51%	↗	4,71%	↗	4,80%
casalinga/stud/altro	19,74%	↘	17,38%	↗	18,19%	↘	15,19%
pensionato/ritirato	19,46%	↘	17,33%	↗	18,50%	↗	18,99%

Tabella 7: La tabella esplica la statistica descrittiva sulle condizioni professionali dei soggetti intervistati.

In considerazione al mercato del lavoro i pesi percentuali tra le varie categorie appaiono pressoché stabili, ci si limita solo ad evidenziare un lieve aumento delle persone ritirate dal lavoro, presumibilmente contestuale all'invecchiamento della popolazione italiana, ed un altro lieve aumento per i soggetti in cerca di lavoro che sebbene sia in notevole recupero dalla situazione emergente nel 2014, ora forse testimonia un po' le sofferenze dovute alla difficolta di accesso a questo settore.

Principali mete di destinazione

Spostando il focus da una valutazione sulla stratificazione sociale verso un topic molto più connesso alla natura vacanziera delle tematiche che si stanno intercorrendo si presentano ora le rilevazioni pertinente alla scelta della destinazione.

8.	Destinazione Italia o estero						
	2014		2018		2019		2020
Italia	80,26%	↘	79,16%	↘	75,83%	↗	92,41%
Estero	19,74%	↗	20,84%	↗	24,17%	↘	7,59%

La tabella 8: La tabella mostra le preferenze di destinazione per i soggetti intervistati.

Nella scelta della destinazione appare evidente l'effetto che l'epidemia ha avuto nel mutare i comportamenti e le scelte delle persone: dal 2014, infatti, si era in presenza di un andamento che vedeva l'apprezzamento delle destinazioni oltre confine, ma le recenti politiche di controllo dei flussi turistici nonché di chiusura delle frontiere hanno portato forzatamente l'Italia ad essere meta di destinazione per tutte quelle persone che altrimenti avrebbero scelto lidi extra-nazionali.
Ciò emerge ancora più chiaramente dalla tabella successiva che esplicita maggiormente quanto detto in precedenza.

9.	Mete di destinazione del viaggio							
	2014		2018		2019		2020	
Piemonte	4,29%	↘	2,61%	↗	3,23%	↘	3,11%	
Valle d'Aosta	0,64%	↗	1,48%	↗	1,79%	↘	0,84%	
Lombardia	6,65%	↘	5,74%	↘	5,32%	↘	4,85%	
Trentino	5,13%	↗	7,51%	↘	7,11%	↗	12,55%	
Veneto	7,77%	↘	7,68%	↘	6,15%	↗	7,33%	
Friuli	1,68%	↗	2,53%	↘	2,27%	↘	1,85%	
Liguria	4,25%	↗	5,15%	↗	5,41%	↘	4,85%	
Emilia-Romagna	9,21%	↘	6,92%	↗	7,72%	↗	10,50%	
Toscana	7,85%	↗	10,08%	↘	9,42%	↗	12,08%	
Umbria	2,76%	↘	1,18%	↘	1,09%	↗	2,11%	
Marche	2,32%	↘	2,19%	↗	3,05%	↗	3,43%	
Lazio	6,17%	↘	5,53%	↘	3,75%	↗	4,27%	
Abruzzo	2,04%	↗	2,19%	↗	2,31%	↗	4,32%	
Molise	0,32%	↘	0,17%	↗	0,31%	↗	0,53%	
Campania	5,01%	↘	2,95%	↗	3,32%	↗	4,64%	
Puglia	5,29%	↘	5,10%	↘	4,62%	↘	4,17%	
Basilicata	0,68%	↘	0,38%	↗	1,48%	↘	0,74%	
Calabria	2,28%	↗	2,32%	↘	1,79%	↗	2,32%	
Sicilia	2,76%	↗	3,50%	↘	2,97%	↗	3,48%	
Sardegna	3,16%	↗	3,96%	↘	2,71%	↗	4,43%	
Austria	0,72%	↗	1,01%	↗	1,09%	↘	0,53%	
Belgio	0,36%	↘	0,08%	↗	0,31%	↘	0,16%	
Danimarca	0,00%	↘	0,00%	↘	0,00%	↘	0,00%	
Finlandia	0,00%	↗	0,13%	↗	0,17%	↘	0,00%	
Francia	3,56%	↘	2,45%	↗	2,53%	↘	1,53%	
Germania	1,52%	↘	1,01%	↗	1,27%	↘	0,58%	
Grecia	0,48%	↗	0,93%	↗	2,05%	↘	0,42%	
Irlanda	0,20%	↘	0,04%	↗	0,22%	↘	0,00%	
Lussemburgo	0,00%	↘	0,00%	↘	0,00%	↘	0,00%	
Olanda	0,12%	↗	0,84%	↘	0,65%	↘	0,11%	
Polonioa	0,12%	↗	0,34%	↘	0,17%	↗	0,37%	
Portogallo	0,52%	↗	0,67%	↘	0,13%	↘	0,11%	
Regno Unito	1,84%	↘	0,97%	↗	1,70%	↘	0,00%	
Rep. Ceca	0,12%	↗	0,55%	↗	0,70%	↘	0,16%	
Slovacchia	0,00%	↘	0,00%	↗	0,04%	↘	0,00%	
Spagna	2,08%	↗	3,04%	↗	4,06%	↘	0,90%	
Svezia	0,00%	↗	0,46%	↘	0,04%	↘	0,00%	
Ungheria	0,12%	↗	0,38%	↗	0,48%	↘	0,11%	
Bulgria	0,00%	↗	0,30%	↘	0,13%	↘	0,00%	
Cipro	0,04%	↗	0,04%	↗	0,04%	↘	0,00%	
Estonia	0,00%	↘	0,00%	↘	0,00%	↘	0,00%	
Latvia	0,00%	↘	0,00%	↘	0,00%	↘	0,00%	
Lituania	0,00%	↘	0,00%	↗	0,13%	↘	0,11%	
Malta	0,12%	↗	0,30%	↘	0,22%	↘	0,00%	
Romania	0,80%	↘	0,38%	↘	0,26%	↘	0,16%	
Slovenia	0,32%	↗	0,42%	↗	0,57%	↘	0,26%	
Croazia	1,16%	↗	1,60%	↗	1,79%	↘	0,84%	
Paesi non EU	2,40%	↘	2,15%	↘	1,53%	↘	0,47%	
Nord America	0,60%	↘	0,55%	↗	0,83%	↘	0,16%	
centro-sud America	0,08%	↗	0,51%	↗	0,52%	↘	0,16%	
Africa	1,48%	↘	0,59%	↗	1,31%	↘	0,32%	
Asia-Oceania	0,96%	↗	1,10%	↗	1,22%	↘	0,16%	

La tabella 9: La tabella mostra le preferenze nelle destinazioni a livello regionale per le mete italiane e a livello paese per quelle oltre confine.

Ripartizioni temporali dei flussi turistici

Dopo aver esaminato dove "gli Italiani" preferiscono passare il loro tempo è di logica consequenziale capire quanto tempo hanno a disposizione da poter passare in vacanza e soprattutto quando decidono di utilizzarlo.
Queste informazioni sono state ricavate all'interno del campionamento rilevazioni sulla durata del viaggio nonché sul relativo mese di inizio, qui percentualmente riportati in tabella.

10.	Mese di inizio viaggio						
	2014		2018		2019		2020
gennaio	4,97%	↘	4,68%	↘	3,93%	↗	7,28%
febbraio	4,57%	↘	3,96%	↗	4,84%	↗	5,96%
marzo	6,05%	↗	7,04%	↘	5,24%	↘	0,63%
aprile	11,05%	↘	8,56%	↗	8,86%	↘	0,00%
maggio	6,61%	↗	6,96%	↘	4,93%	↘	0,63%
giugno	9,33%	↘	9,28%	↗	11,26%	↘	11,13%
luglio	11,49%	↗	15,10%	↘	14,35%	↗	19,57%
agosto	21,43%	↘	21,00%	↗	22,51%	↗	35,34%
settembre	6,33%	↗	7,93%	↘	6,50%	↗	12,87%
ottobre	5,33%	↘	3,92%	↗	6,06%	↘	3,06%
novembre	4,29%	↘	3,54%	↗	4,23%	↘	0,32%
dicembre	8,57%	↘	8,01%	↘	7,29%	↘	3,22%

Tabella 10: La tabella indica le frequenze relative della distribuzione dei mesi di inizio vacanza.

Prescindendo dal pesare le informazioni ottenute durante il corso del 2020 come gli anni precedenti, in contesto di normalità, emerge la presenza di forti e ricorrenti stagionalità a caratterizzazione dei flussi turistici; questi appaiono, infatti, concentrarsi nei periodi estivi nonché quelli densi di ricorrenze come dicembre e gennaio.
In un'ottica intertemporale appare un lieve apprezzamento per le vacanze sostenute nei mesi estivi a discapito delle stagioni più fredde probabilmente dovuto all'innalzamento delle temperature che rende più meno accessibili località sciistiche o comunque correlate all'esigenza di temperature più fredde.
Per quanto riguarda la durata del soggiorno le rilevazioni evidenziano una notevole concentrazione di vacanze con durata uguale o inferiore alla settimana, tuttavia, si

considera pressoché stabile la quantificazione di vacanze di durata superiore che, come ci si potrebbe aspettare, presentano una frequenza sempre più esigua al crescere dei giorni. Una breve considerazione riguardo all'anno 2020 emerge dal fatto che nonostante le restrizioni imposte si sono mantenute le stagionalità emerse negli anni precedenti e che, sebbene si siano ridotti i soggiorni di breve durata, c'è stato un condiviso apprezzamento per la vacanza più classica per l'immaginario comune: ovvero quella dalla durata canonica di una settimana.

11.	Durata del viaggio espressa in giorni							
	2014		**2018**		**2019**		**2020**	
1	17,06%	↘	15,39%	↗	15,84%	↘	13,03%	↘
2	18,34%	↗	19,19%	↘	17,63%	↘	16,93%	↘
3	13,74%	↗	13,79%	↘	13,22%	↘	12,76%	↘
4	7,89%	↗	8,31%	↗	9,25%	↘	8,91%	↘
5	6,61%	↘	6,41%	↘	5,76%	↘	5,54%	↘
6	6,29%	↘	6,24%	↘	5,76%	↘	5,17%	↘
7	9,45%	↗	10,12%	↗	10,73%	↗	14,93%	↗
8	3,40%	↘	2,74%	↘	2,71%	↗	3,06%	↗
9	1,88%	↗	2,32%	↘	1,83%	↘	1,37%	↘
10	3,92%	↘	3,54%	↗	3,58%	↗	4,01%	↗
11	0,56%	↗	0,72%	↗	1,09%	↘	0,84%	↘
12	1,08%	↘	0,97%	↘	0,87%	↗	1,11%	↗
13	0,84%	↗	0,93%	↘	0,57%	↗	0,84%	↗
14	2,60%	↗	3,46%	↗	3,97%	↘	3,22%	↘
15	1,88%	↘	1,64%	↗	2,92%	↘	2,90%	↘
16	0,76%	↗	0,93%	↘	0,35%	↗	0,58%	↗
17	0,32%	↗	0,38%	↗	0,70%	↘	0,47%	↘
18	0,12%	↗	0,13%	↗	0,39%	↗	0,42%	↗
19	0,20%	↘	0,08%	↘	0,00%	↗	0,37%	↗
20	0,72%	↗	0,93%	↗	1,22%	↘	0,69%	↘
21	0,08%	↗	0,63%	↘	0,26%	↗	0,74%	↗
22	0,08%	↗	0,13%	↗	0,35%	↗	0,53%	↗
23	0,16%	↗	0,17%	↘	0,09%	↗	0,21%	↗
24	0,00%	↗	0,30%	↘	0,13%	↘	0,00%	↘
25	0,36%	↘	0,17%	↗	0,17%	↘	0,11%	↘
26	0,08%	↘	0,00%	↘	0,00%	↗	0,11%	↗
27	0,04%	↘	0,00%	↗	0,09%	↘	0,00%	↘
28	0,20%	↘	0,13%	↘	0,00%	↗	0,32%	↗
29	0,08%	↘	0,04%	↗	0,04%	↗	0,11%	↗
30	0,64%	↘	0,08%	↗	0,22%	↗	0,53%	↗
31	0,00%	↘	0,00%	↘	0,00%	↗	0,11%	↗
32	0,04%	↗	0,08%	↘	-	↔	-	
33	-	↔	-	↗	0,09%	↘	-	
35	-	↗	0,04%	↗	0,09%	↘	-	
37	-	↔	-	↗	0,04%	↘	-	
40	0,20%	↘	-	↗	0,04%	↗	0,05%	
41	0,08%	↘	-	↔	-	↔	-	
57	0,08%	↘	-	↔	-	↔	-	
60	0,08%	↘	-	↔	-	↗	0,05%	
63	0,08%	↘	-	↔	-	↔	-	
80	0,04%	↘	-	↔	-	↔	-	

Tabella 11. La tabella delinea le frequenze relative per la durata della vacanza pesate sul numero di giorni di ferie indicati dai soggetti intervistati.

Principali criteri di scelta per la vacanza

Questa macrocategoria comprende l'indicazione di tutte quelle scelte che caratterizzano l'esperienza turistica; quindi, si andrà ad analizzare ad esempio la scelta dell'alloggio ma anche il mezzo di trasporto necessario per poter arrivare alla destinazione di residenza desiderata.

12.	Mezzo di trasporto utilizzato per il raggiungimento della meta turistica							
	2014		2018		2019		2020	
aereo	15,82%	↗	17,00%	↗	20,07%	↘	7,12%	
treno	9,45%	↘	7,68%	↘	6,98%	↘	4,48%	
nave, battello	2,72%	↗	4,01%	↘	2,14%	↗	2,27%	
auto a noleggio	1,04%	↗	1,60%	↘	1,35%	↗	1,95%	
auto propria	61,03%	↗	61,45%	↘	59,08%	↗	76,74%	
pullman turistico	4,04%	↘	3,80%	↗	5,06%	↘	1,16%	
pullman di linea	1,72%	↘	0,72%	↗	0,74%	↘	0,47%	
camper, caravan	3,04%	↘	2,40%	↗	3,53%	↗	3,74%	
moto, scooter	0,40%	↗	0,55%	↗	0,65%	↗	1,42%	
altro	0,72%	↗	0,80%	↘	0,39%	↗	0,63%	

Tabella 12: La tabella mostra le frequenze di utilizzo dei mezzi necessari per il raggiungimento delle mete vacanziere.

Nella scelta del mezzo di trasporto, il passare degli anni han portato alla luce la predilezione per l'utilizzo di soluzioni di spostamento mediante il ricorso alle sempre più diffuse ed economiche offerte delle compagnie aree, le quali hanno totalmente sostituito il trasporto su ruote che comunque presenta un lieve rialzo, sostenuto anche questo da multinazionali del settore come Flixbus.

Con il 2020 i trend giungono contestualmente ad interruzione portando conseguentemente ad un notevole incremento del trasporto su ruote che garantisce la massima flessibilità nonché il rispetto di forme di distanziamento sociale.

13.	Principali luoghi di destinazione per la vacanza						
	2014		2018		2019		2020
mare	28,88%	↗	33,91%	↗	35,01%	↗	38,14%
crociera	0,44%	↗	0,88%	↘	0,56%	↘	0,04%
montagna	16,67%	↗	18,30%	↘	17,16%	↗	22,71%
città	39,56%	↘	32,54%	↗	34,91%	↘	24,81%
campagna	9,86%	↗	9,93%	↘	8,81%	↗	12,13%
altro	4,59%	↘	4,45%	↘	3,55%	↘	2,18%

Tabella 13: La tabella mostra i principali luoghi di destinazioni aggregati per tipologia.

Dal punto di viste delle categorie di destinazione, la fanno da padrone le mete balneari e cittadine che si mantengono stabili ai vertici delle preferenze; l'effetto pandemico non ha stravolto troppo i pesi dei partizionamenti se non per una diminuzione delle mete urbane in favore delle destinazioni montane o di campagna tipiche del patrimonio paesaggistico-culturale italiano.

14.	Tipologia di attività principale del viaggio						
	2014		2018		2019		2020
scoperta del territorio	-		42,79%	↘	24,96%	↘	22,18%
cura e arricchimento	-		2,99%	↗	3,36%	↗	3,40%
intrattenimento, svag	-		54,22%	↗	57,46%	↗	74,42%

Tabella 14: La tabella mette in evidenza le preferenze relative in merito agli scopi della vacanza.

La tabella 14., a partire dall'anno 2018, permette di ricavare informazioni in merito alle attività svolte dai soggetti intervistati durante il periodo vacanziero; non appaiono particolari evidenze, i viaggi per cura e arricchimento personali si mantengono stabili tra le rilevazioni, l'unica variazioni si ha per l'aumento della categoria connessa all'intrattenimento e lo svago, segno che emerge la generale tendenza a preferire vacanze in cui sono importanti le attività che si andranno a svolgere

Si precisa che per le difformità presenti nelle risposte indicate nei vari records, le quali dal 2019 in poi prevedono un format d'inchiesta differente per la fase di campionamento, il grafico riportato è stato ricavato da un ulteriore processo di pulizia del campione che verrà affrontato e descritto successivamente.

15.	Alloggio						
	2014		2018		2019		2020
albergo/motel/pensic	30,52%	↗	44,16%	↘	42,80%	↘	38,03%
residenza di cura fisic	0,60%	↘	0,25%	↗	0,48%	↘	0,16%
campo lavoro e vacar	0,12%	↘	0,04%	↗	0,04%	↗	0,11%
mezzo pubblico di tra	0,04%	↗	0,42%	↗	0,70%	↘	0,11%
centro congressi e co	0,04%	↘	0,00%	↘	0,00%	↘	0,00%
villaggio vacanza	1,72%	↗	2,66%	↘	2,18%	↗	3,38%
campeggio	4,04%	↗	5,02%	↗	5,28%	↗	8,02%
marina	0,32%	↘	0,25%	↘	0,04%	↗	0,05%
istituto religioso	0,56%	↗	0,67%	↘	0,57%	↘	0,11%
altra struttura colletti	1,44%	↘	1,27%	↗	1,31%	↘	0,63%
stanza in affitto	0,40%	↗	0,72%	↗	0,92%	↘	0,53%
abitazione in affitto	5,05%	↗	10,12%	↗	11,17%	↘	10,65%
abitazione propria o r	8,17%	↗	9,19%	↘	7,42%	↗	9,28%
abitaz. parenti o amic	36,64%	↘	14,21%	↗	14,92%	↘	14,29%
barca in sito non orga	0,52%	↘	0,08%	↗	0,09%	↘	0,05%
altro tipo di sistemazi	2,76%	↘	1,77%	↗	1,96%	↘	1,69%
agriturismo	1,44%	↘	1,10%	↗	1,53%	↗	2,85%
B&B	5,61%	↗	8,06%	↗	8,60%	↗	10,07%

Tabella 15: La tabella mostra le preferenze relative nella scelta degli alloggi da parte dei soggetti intervistati.

A conclusione, si giunge a mostrare le evidenze ottenute in considerazione della scelta di alloggio: lo scenario ha fatto emergere negli anni una notevole predilezione per soluzioni di alloggio connesse a strutture ricettizie di tipo alberghiero e affini. Il resto delle voci si mantiene fluttuante nel tempo senza indicizzare particolari orientamenti eccezion fatta per l'utilizzo di abitazioni appartenenti a parenti o ad amici, la categoria, per l'appunto, mostra una notevole flessione sul medio periodo passando da soluzione più popolare a mera alternativa della prenotazione alberghiera.

La consueta considerazione sull'anno 2020 si fonda sul mantenimento, pressoché inalterato dei vari equilibri emersi in precedenza con solo modesti incrementi per il settore dei B&B e dei campeggi sommariamente bilanciati da riduzione nel settore alberghiero, risaputamene in sofferenza soprattutto all'inizio del periodo pandemico.

Iter organizzativo e adesioni alla vacanza

Di seguito vengono presentate le informazioni ottenute dalle valutazioni sulle scelte organizzative in merito alle prenotazioni di viaggio e alloggio poste in essere dal soggetto intervistato, nonché, dall'eventualità, che questo abbia predisposto la vacanza affinché possa essere di fruizione ad altri membri della propria famiglia.

16.	Organizzazzione alloggio							
	2014		2018		2019		2020	
prenotazione diretta	32,28%	↗	52,76%	↘	28,36%	↗	49,37%	
prenot. In agenzia/to	9,25%	↗	11,77%	↗	37,48%	↘	16,35%	
nessuna prenot.	58,31%	↘	32,94%	↘	30,37%	↗	31,86%	
non sa/non risp.	0,16%	↗	2,53%	↗	3,80%	↘	2,43%	

17.	Organizzazione trasporto							
	2014		2018		2019		2020	
prenotazione diretta	16,94%	↗	22,48%	↘	21,86%	↘	11,02%	
prenot. In agenzia/to	8,45%	↘	8,22%	↗	11,74%	↘	4,54%	
nessuna prenot.	74,45%	↘	68,71%	↘	65,71%	↗	84,44%	
non sa/non risp.	0,16%	↗	0,59%	↗	0,70%	↘	0,00%	

Tabelle 16 e 17: Le tabelle evidenziano rispettivamente: i criteri relativi per l'organizzazione delle scelte di alloggio e i criteri relative all'organizzazione dei mezzi di trasporto necessari per la vacanza.

Limitatamente alle modalità di prenotazione alloggio negli anni si è notevolmente ridotto il numero di soggetti che non effettua una prenotazione.
Dal 2014 al 2019 gli intervistati hanno evidenziato una crescente popolarità nel rivolgersi ad agenzie tour operator a discapito delle prenotazioni dirette; questo scenario si ribalta quasi nel corso del 2020, probabilmente in relazione alla chiusura di molte agenzie dovuta al blocco, quasi totale, del settore turistico.

18.	Utilizzo di internet per la prenotazione dell'alloggio							
	2014		2018		2019		2020	
Si	24,15%	↗	47,11%	↘	39,14%	↗	43,35%	
No	17,26%	↘	15,48%	↗	22,95%	↘	20,89%	
Non sa	58,59%	↘	37,41%	↗	37,91%	↘	35,76%	

19.	Utilizzo di internet per prenotare il trasporto							
	2014		2018		2019		2020	
Si	14,98%	↗	21,97%	↗	22,64%	↘	10,86%	
No	10,25%	↘	7,55%	↗	9,95%	↘	4,59%	
Non sa	74,77%	↘	70,48%	↘	67,41%	↗	84,55%	

Tabelle 18 e 19: Le restituiscono le percentuali di soggetti che hanno utilizzato internet per la prenotazione del proprio alloggio o mezzo di trasporto.

A chiudere la tematica riguardante le prenotazioni si presentano le tabelle 18 e 19 che riassumono i dati relativi all'utilizzo di internet da parte del soggetto prenotante; sebbene i dati emersi soffrano di una notevole quantità di risposte non chiare in merito alla questione poiché molto spesso il soggetto intervistato non conosce in che modo è stata effettuata la prenotazione.
In entrambi i casi, ad esempio, è riscontrabile un andamento concorde per le prenotazioni dirette dell'alloggio e l'utilizzo di internet; lo stesso vale in merito agli andamenti delle risposte "prenotazione in agenzia/tour operator" e "No" in riguardo ai trasporti, però, in questo caso, diversamente dal precedente l'interpretazione del comune andamento appare più complessa.
Per quanto, invece, riguarda l'estensione del viaggio più famigliari dalle sottostanti tabelle 20 e 21 si riscontra l'ormai, sempre più, consolidata e diffusa abitudine degli italiani a viaggiare principalmente con la famiglia; negli anni si evidenzia una riduzione dei viaggi di coppia mentre si mantengono solide nel corso del tempo le famiglie di vacanzieri con uno o due figli a carico.

20.	Partecipano al viaggio altri membri della famiglia							
	2014		**2018**		**2019**		**2020**	
Si	65,40%	↗	71,45%	↘	70,24%	↗	74,47%	
No	34,60%	↘	28,55%	↗	29,76%	↘	25,53%	

21.	Numero di membri famigliari partecipanti al viaggio							
	2014		**2018**		**2019**		**2020**	
1	34,60%	↘	28,55%	↗	29,76%	↘	25,53%	
2	37,40%	↗	39,06%	↗	40,01%	↘	39,77%	
3	14,86%	↗	15,86%	↘	15,58%	↗	18,35%	
4	10,73%	↗	14,38%	↘	12,57%	↗	13,92%	
5	2,40%	↘	2,15%	↘	2,09%	↗	2,43%	

Tabelle 20 e 21: Le tabelle evidenziano le percentuali di soggetti che hanno sostenuto una vacanza in compagnia di altri membri della propria famiglia e le percentuali con cui questi si dividono in base ad un'indicazione di numerosità.

3.4. Alcune semplificazioni necessarie per la stima del modello di rete Bayesiana

Nonostante i passaggi precedentemente effettuati per eliminare le variabili superflue e rendere la base campionaria "consistent", l'analisi descrittiva appena compiuta ha reso evidente e necessaria l'applicazione di ulteriori semplificazioni sull'insieme dei dati a causa di numerosi outliers nelle rilevazioni e modalità che potrebbero essere combinate tra loro.

Sulla base di ciò le rilevazioni riguardanti le destinazioni delle vacanze sono state semplificate rimuovendo la gran parte delle destinazioni estere; come si può notare dalla tabella 9., la quasi totalità delle mete all'infuori dei confini italiani evidenzia delle percentuali estremamente ridotte, per cui si è deciso di mantenere solo le destinazioni: Francia, Germania, Grecia, Spagna, Croazia e Inghilterra; le quali da sole rappresentato la quasi totalità di questa categoria di destinazioni.

A giustificazione di quanto fatto sta' il desiderio di voler dare comunque rappresentanza alle località estere, che, sebbene in numerosità molto ridotta, appaiono comunque significative e contribuiscono ulteriormente ad ampliare la stratificazione informativa delle osservazioni raccolte, mantenendo, nel contempo, saldi i principi di semplificazione.

Analogamente con quanto fatto per le destinazioni, sono state rimossi tutti i records relativi alle vacanze di durata superiore ai ventun giorni e quelli riferibili alla variabile "tipo crociera"; la scelta di operare in questo modo deriva, innanzitutto, dai dati presenti nella tabella 11. che denotano, anche qui, percentuali estremamente basse per le vacanze superiori alle tre settimane e, poi, per l'esiguo numero di vacanze svolte in crociera, tanto da risultare irrilevanti se confrontate con le altre categorie tipologiche (si veda tabella 13.).

Inoltre, sono state eseguite delle operazioni di rielaborazione dei records che hanno permesso di riaggregare alcune modalità di risposta dai tratti quasi sovrapponibili, al fine di avere delle categorizzazioni più compatte e senza outliers.

Questo ha previsto per la variabile relativa alle scelte di alloggio una razionalizzazione logica delle numerose voci presenti in sole tre modalità: Strutture ricettizie, Locazione in affitto o B&B e Alloggio privato.

Le prime aggregano le rilevazioni delle precedenti voci: Alberghi, motel e pensioni; a cui si aggiungono per pertinenza quelle delle residenze di cura, dei villaggi vacanza, dei campeggi, degli istituti religiosi e altre strutture collettive.

La seconda categoria, più semplicemente, raggruppa tutti gli alloggi per cui si paga un affitto (si associa a questa categoria anche il B&B che per le sue dinamiche risulta simile negli approcci all'affitto di una stanza nonché per la sua connotazione meno "commerciale" più pertinente, invece, all'attività alberghiera vera e propria) perciò vi ricadono le osservazioni pertinenti gli alloggi in: stanze in affitto, abitazioni in affitto e B&B.

Con l'ultima voce, infine, si fanno convergere tutte le soluzioni relative ad abitazioni di proprietà quindi per le quali non si prevede un esborso in denaro, il che comprende: le abitazioni proprie o in multiproprietà, le abitazioni di parenti o amici e altre tipologie di alloggi privati.

Al termine di questo processo sugli alloggi restano, tuttavia, non allocate le seguenti voci: campo da lavoro, mezzo pubblico, centro congressi, marina, barca in sito non organizzato e agriturismo. In ragion dell'esigua numerosità di queste e per la difficoltà di attribuzione logica ad una delle tre macrocategorie proposte, state eliminate dalla base campionaria.

Segue, a quanto fatto per gli alloggi, una riaggregazione anche per i mezzi di trasporto che, nella nuova proposizione, mantengono inalterate le voci: aereo, treno e nave; aggiungendo a queste le categorie: su ruote autonomo e pullman.

La pima modalità raggruppa tutte quelle soluzioni di trasporto su ruote dipendenti "agire" del soggetto intervistato, quindi: l'utilizzo di un'auto propria, un auto a noleggio, moto o scooter, camper e altre soluzioni.

Logicamente opposta risulta la categoria pullman che specifica le soluzioni di trasporto su ruote in cui il soggetto intervistato non ha la necessità di dover guidare ma il suo agire si limita solo all'acquisto della tratta; di conseguenza rispecchiano la categoria le precedenti voci: pullman turistico e pullman di linea.

L'ultimo insieme di voci a subire una riaggregazione è quello esplicitante le attività del soggiorno. rispetto alle variabili precedenti tra gli anni di rilevazione non si è mantenuto il medesimo schema di domande per le interviste e questo ha fatto si che anche le modalità assunte per questa variabile fossero diverse per alcuni degli anni considerati.

Sebbene queste difficoltà di incoerenza tra gli output di rilevazione, si è cercato di restituire una variabile con modalità aggregate che potesse adattarsi alle differenze emerse: per questo si sono individuate tre macro categorie ad ampio spettro, allo scopo allocare tutte le osservazioni, ovvero: cura e arricchimento personale, scoperta del territorio e intrattenimento e svago.

Le analisi ottenute al termine di queste elaborazioni sulle modalità di alcune variabili, presenta solo alcune piccole variazioni rispetto ai vari equilibri percentuali esaminati precedentemente e quindi restano valide le conclusioni tratte sull'analisi effettuata a prtire dalle variabili originali.

Analisi della capacità di spesa

Passando ora ad un abito più strettamente economico, ad esclusione dell'anno 2014 in cui queste indicazione non erano ancora presenti, lo studio delle variabili relative alle informazioni sulla spesa sostenuta dai soggetti intervistati permette di formulare delle idee di base sulla capacità di spesa per le vacanze.
Nei dati analizzati sono presenti due variabili che indicano rispettivamente la spesa media giornaliera per una vacanza e il relativo esborso complessivo, ottenibile dal prodotto tra spesa giornaliera e durata del viaggio, sostenuti da chi ha risposto al questionario; queste informazioni permettono di capire, soprattutto in un'ottica intertemporale, l'attrattività di questo settore nonché la quantità di risorse finanziarie che i soggetti intervistati possono dedicare alle vacanze.

	2018		2019		2020
spesa media complesiva	246,71	↗	282,69	↗	399,31

Tabella 22: la tabella mostra le spese medie complessive sostenute dai soggetti intervistati per effettuare le loro vacanze.

Da un confronto sui dati a disposizione emerge un andamento medio di spesa in costante crescita, sostanzialmente, almeno basandosi sul puro dato, appare più costoso andare in vacanza nel 2020 rispetto al 2018; tuttavia una comprensione maggior su questi dati si ottiene andando a confrontarli con la durata media delle vacanze, in quanto la durata influenza direttamente l'ammontare esborsato.

	2018		2019		2020
durata media della vacanza	5,38	↗	5,61	↗	6,02
spesa giornaliera media	45,85	↗	50,43	↗	66,30

Tabella 23: La tabella mostra le spese medie giornaliere confrontate con la durata media delle vacanze effettuate dai soggetti intervistati.

Quanto detto si esplicita molto chiaramente nella tabella 23: durata della vacanza e spesa media giornaliera seguono di pari passo l'andamento del costo totale, tuttavia, l'aumento di prezzi è solo parzialmente da imputarsi all'aumento di giorni in vacanza, il costo cresce molto più rapidamente del tempo in ferie, quindi si può effettivamente suggerire che il settore abbia subito una forte inflazione nel corso di questi ultimi anni.
Ciò risulta ulteriormente visibile da un confronto sulle distribuzioni normalizzate delle spese complessive (Grafico 3.1 e Grafico 3.2); prendendo in riferimento solo gli anni 2018 e 2020, appare chiaro che la curva di densità nel tempo abbia subito una notevole trasformazione riducendo la propria pendenza.

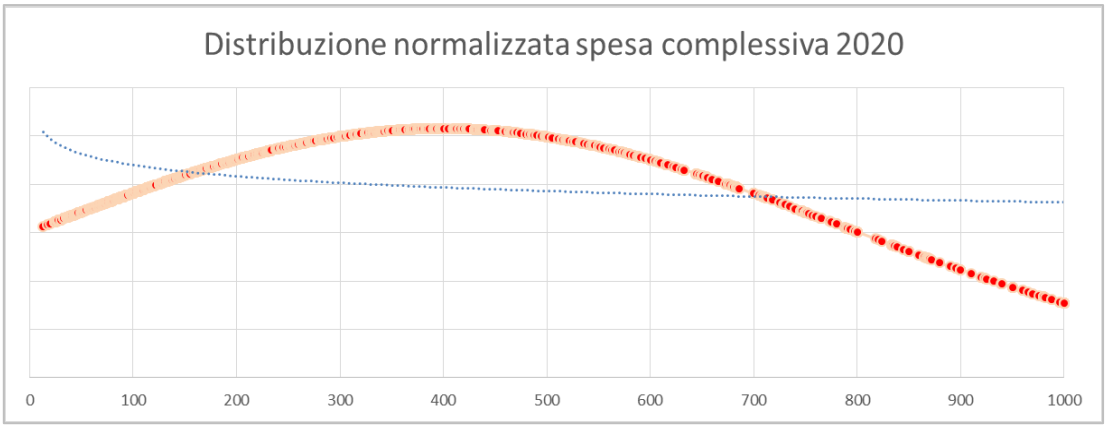


Grafico 3.1: Distribuzione normalizzata della variabile spesa complessiva per l'anno 2020

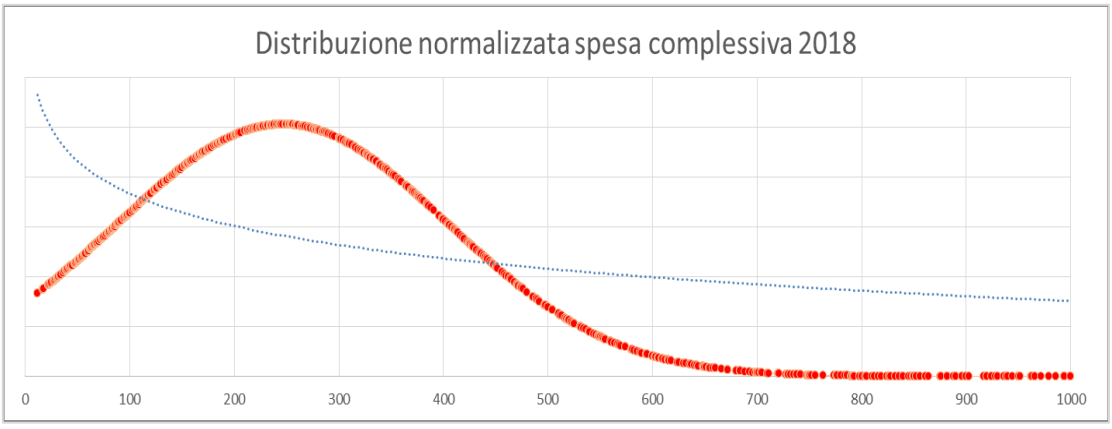


Grafico 3.2: Distribuzione normalizzata della variabile spesa complessiva per l'anno 2018

3.5 Le reti Bayesiane per lo studio delle relazioni sulle caratteristiche delle vacanze

Giunti a questo punto si procederà ora con la stima dei modello di reti Bayesiane per gli anni presi in considerazione. Per far ciò si utilizzerà la libreria bnlearn[20] del software R. Tale libreria mette a disposizione numerosi algoritmi di apprendimento della struttura, sia della categoria costraint based sia score based (nonché misti); perciò, dopo aver testato ogni singolo algoritmo, effettuato un confronto delle reti create e una valutazione del relativo punteggio, tramite una funzione di score ad approccio BIC, è stato possibile identificare l'algoritmo hill Climbing come più performante e adatto all'individuazione del network.
Selezionato l'algoritmo più adatto, la rappresentazione grafica della rete permette di poter esprimere le prime considerazioni sul fenomeno e di fare, anche, dei confronti sul suo andamento negli anni.
Prima che ci si addentri nella discussione, è bene fare alcune premesse sull'anno 2014, poiché, non essendo uniforme con i successivi anni, ha mostrato alcune differenze nella struttura appresa: sebbene ad un micro livello alcune sezioni della rete mantengono stabili negli anni i loro legami (ci si riferisce ad esempio all'insieme delle variabili anagrafiche o di tipologia della destinazione), si ritiene che l'assenza di alcune variabili, come quelle relative alla spesa sostenuta per il viaggio (e marginalmente anche quella che identifica l'attività principale della vacanza), abbia alterato, rispetto agli anni successivi, la determinazione dei nodi e dei legami per un confronto coerente.

Anno 2020:

La figura 3.1 mostra gli esiti del processo descritto nel paragrafo precedente; la rete Bayesiana, così prodotta, mostra per l'anno 2020 tutte le connessioni presenti tra le variabili del dataset, inoltre, la posizione relativa di queste, nelle varie nodosità del network, permette di dare una sorta di ordine al processo turistico, che altrimenti non si sarebbe potuto cogliere.

[20] Si rimanda al sito web ufficiale di questa libreria: https://www.bnlearn.com/

L'analisi di questa rete segue la sua struttura topologica, ma, già ad un primo impatto, è possibile individuare tre aree spaziali, suddivise in base alle informazioni ricavabili dalle variabili.

Queste aree raggruppano al loro interno, rispettivamente: le nodosità riguardanti la sfera anagrafica del viaggiatore, quelle pertinenti alla scelta della tipologia di destinazione dove andare a svolgere la vacanza e, infine, quelle connesse all'iter organizzato per l'alloggio e i mezzi di trasporto; questa suddivisione della rete, come si vedrà in seguito, non è esclusiva per l'anno 2020, bensì, si riscontrerà, anche, nelle reti degli anni precedenti; segno di una stabilità del fenomeno turistico, ma, soprattutto, di una coerenza analitica del modello Bayesiano.

Tornando allo studio di quanto emerso per il 2020, il network presenta come nodo radice la ripartizione geografica di residenza, la quale influenza direttamente la destinazione regionale del viaggio[21]; risulta possibile, quindi, già dall'analisi del primo arco, una di differenziazione geografica per i soggetti intervistati.

La variabile destinazione regionale, si dirama, poi, in due sentieri diversi: uno che sfocerà nell'area spaziale pertinente alla tipologia della meta scelta, mentre, l'altro andrà a collegarsi con l'area spaziale in cui sono presenti le variabili connesse alla sfera organizzativa del viaggio.

La parte di network che identifica l'area pertinente alla tipologia della destinazione, è individuabile nella parte sinistra della rete; essa comprende le variabili: "tipo mare", "tipo montagna", "tipo città", "tipo campagna" e "tipo altro".

Limitatamente a questa zona del network si può notare come le destinazioni di tipo marittimo abbiano una posizione di vertice rispetto alle altre; tale disposizione potrebbe essere interpretata alla luce di quanto emerso dalle statistiche descrittive, le quali denotano una trasversale preferenza per la vacanza in siti balneari.

Concentrandosi sul nodo tipo mare emerge, anche, un legame di connessione diretta tra questa variabile con il mese d'inizio e la durata della vacanza; ciò segnala una forte dipendenza alla stagionalità per questo tipo di destinazione e anche una durata maggiore del soggiorno. Di conseguenza si può dire che, a grandi linee, gli Italiani preferiscono località marittime per la vacanza principale, mentre le ferie in altre destinazioni hanno natura marginale e con durate inferiori, a riprova della posizione inferioire delle altre variabili tipologiche rispetto al nodo "TIPOMARE".

[21] Si rimanda all'appendice dove sarà riportata una descrizione di tutte le variabili

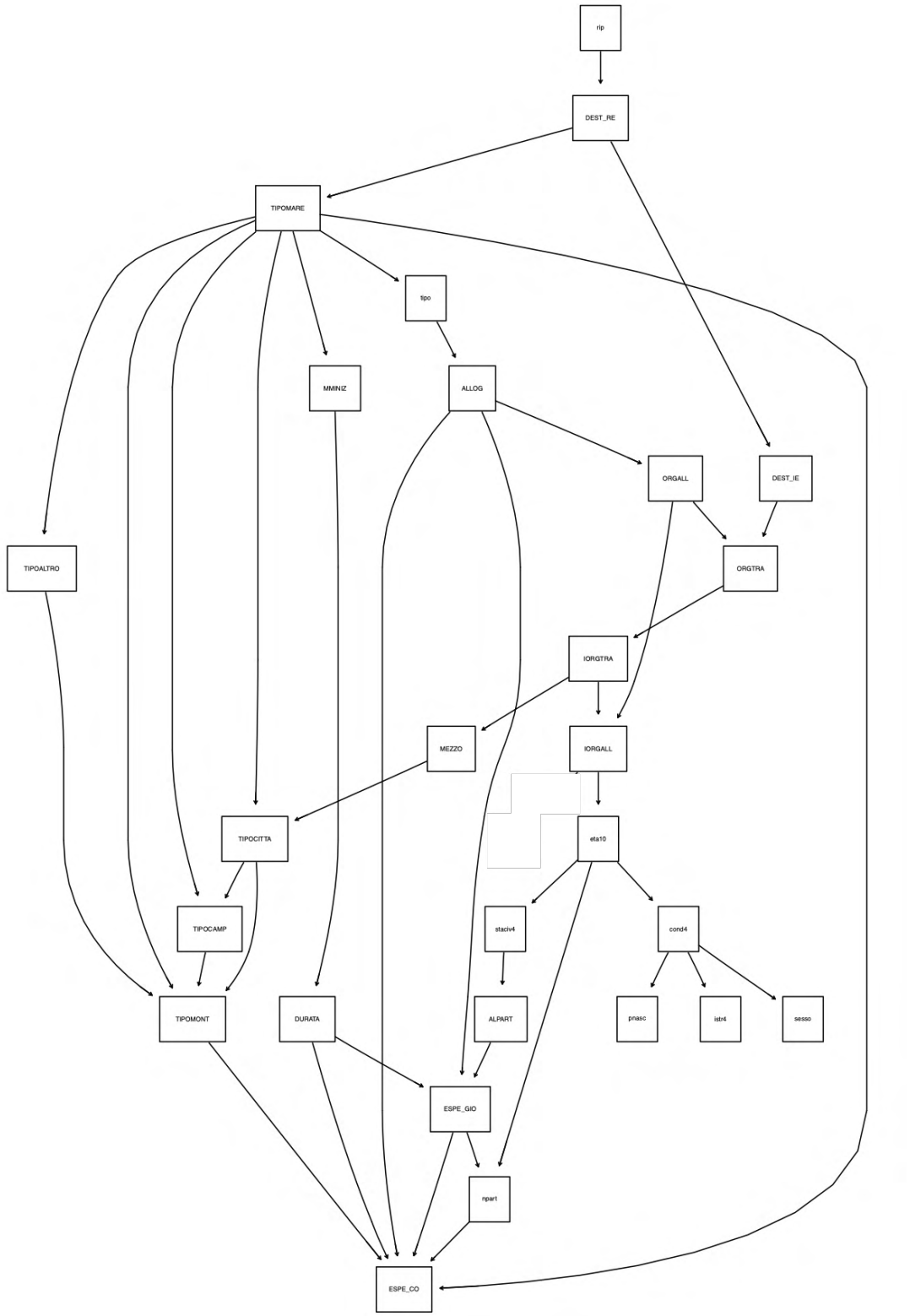


Figura 3.1: Rete Bayesiana anno 2020

Per quanto riguarda la sfera organizzativa essa comprende, invece, tutti i nodi che partendo dalla viabile “alloggio” conducono fino alla definizione del mezzo e alle variabili di natura anagrafica.

Quest'area del network, che per tutti gli anni in rilevazione ricomprende i medesimi nodes, mostra il primato della scelta della dimora (si noti come questa sia in relazione indiretta, mediante la variabile tipo, con il nodo "TIPOMARE", segno ulteriore della centralità delle destinazioni marittime nel contesto vacanziero) sul metodo per poterla scegliere, in altri termini, per l'anno 2020 prima si individua la sistemazione, poi ci si organizza per averla.

La sezione del network in cui si ritrovano le variabili di natura anagrafica mostra il nodo "eta10" in posizione apicale; questo nodo presenta come suo genitore la variabile "iorgall", pertinente all'utilizzo di internet per la prenotazione dell'alloggio, segno di una possibile correlazione tra età del soggetto intervistato con l'utilizzo di determinati canali prenotativi.

La variabile età, nel diramarsi tra le altre variabili di contesto, mostra dei sentieri intuitivamente logici, non facendo emergere nessuna particolare considerazione se non forse in merito alla variabile della condizione lavorativa, la quale influenza: il sesso, l'istruzione e il paese di nascita; sebbene queste influenze possano avere limitata utilità in un contesto turistico, la loro conformazione può risultare rilevante nei termini di un'analisi di stratificazione sociale.

Le variabili di spesa, infine, fanno da collettore di fondo a tutto il processo; su di esse ricadono tutte le sezioni esaminate; particolarmente forti, data la presenza di connessioni più dirette, sono i legami con: la tipologia delle destinazioni, le variabili di durata della vacanza e quelle inferiori dell'area anagrafica che esplicitano la partecipazione o meno di altri membri della famiglia alla vacanza.

La disposizione delle variabili, i legami emersi tra esse, nonché, le associazioni logiche che hanno mostrato saranno tutte informazioni rilevanti per lo sviluppo di un pensiero critico sul fenomeno analizzato; in quanto l'esplicitazione visiva delle relazioni di probabilità condizionata permette di avere una maggior consapevolezza delle reali informazioni presenti in un dataset.

Anno 2019[22]:

Per quest'anno, e anche per i precedenti, la trattazione prescinderà dal fornire una completa disamina della rete come fatto per il 2020; questa scelta è stata fatta in relazione

[22] Si rimanda all'appendice per una visione completa della rete Bayesiana di quest'anno

all’omogeneità delle strutture Bayesiane, che di anno in anno, si presentano estremamente simili; perciò, sulla base di questo, la trattazione si limiterà ad evidenziare solo peculiarità o dettagli degni di nota.

Un confronto tra le reti del 2019 e del 2020 mostra delle strutture pressoché sovrapponibili, la presenza di alcune differenze tra le diramazioni degli archi risulta trascurabile; nel network, infatti, le medesime tre aree spaziali, già evidenziate nell’anno 2020, rispecchiano l’anno successivo nel fluire dei loro processi relazionali; in termini macro, rimane, quindi, inalterato il comportamento delle variabili espresso dalla struttura.

In questo contesto due sono le uniche differenze di cui far menzione: la prima, più marginale, riguarda una maggior rappresentatività per la variabile “TIPOMONT” che si mostra come scelta autonoma e distinta dalla variabile “TIPOMARE”, l’altra, invece, riguarda l’orientamento delle variabili presenti nell’area demandata all’organizzazione della vacanza.

Quest’area rispetto all’anno 2020 mostra, pur mantenendo inalterato il gruppo delle variabili di contesto, un’inversione del trend precedentemente esaminato; per cui l’alloggio non è più in posizione apicale, bensì, si ritrova come nodo di termine, ponendo come prima scelta del processo organizzativo della vacanza la scelta del mezzo (Figura 3.2).

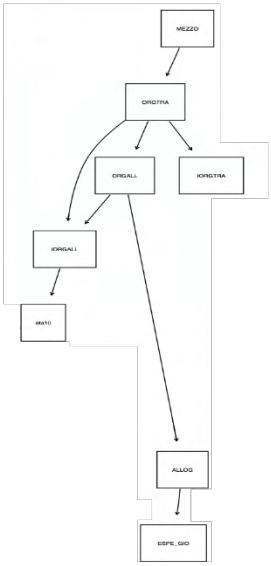


Figura 3.2: La figura rappresenta le variabili pertinenti all’organizzazione della vacanza

Un motivo di questa inversione è individuabile nella pandemia, la quale ha sconvolto il settore turistico comportando la chiusura di molti tour operator e lasciando le strutture ricettive non più a pieno regime; in questo scenario l'alloggio si presenta come discriminante principale, in quanto in mancanza di questo la vacanza non sarebbe altresì realizzabile.

Anno 2018[23]:

Anche per il network di quest'anno valgono le premesse effettuate per il 2019; la strutturazione non presenta anomalie, i legami relazionali si mantengono saldi e coesi tra con rispettive aree topologiche.
La rete generata dai dati del 2018 risulta quasi sovrapponibile a quella dell'anno successivo, si evidenzia solo una mancata indipendenza della variabile "TIPOMONT" che si ripresenta come nodo figlio della controparte "tipo mare"

Anno 2014:

La valutazione di quest'anno, rispetto alle precedenti, acquista particolare valore per la conformazione del dataset di riferimento.
A differenza degli anni successivi, le rilevazioni effettuate nel 2014 non prevedevano le variabili di spesa e quella indicante la tipologia di attività principale svolta durante la vacanza.
Queste discrepanze hanno avuto delle ricadute sulla strutturazione della rete Bayesiana (Figura 3.3); tuttavia, si vuole far notare che queste differenze riguardano esclusivamente la disposizione, di vertice o di fondo, delle medesime aree topologiche descritte nei capitoli precedenti.
In altri termini, l'assenza di queste variabili chiave, ha fatto sì che si identificassero diversi nodi radice e foglia per il delineamento della macrostruttura di rete.

[23] Si rimanda al network completo in appendice

In questo modo la rete dirama le sue relazioni partendo dalla definizione delle variabili di contesto anagrafico, le quali a loro volta ricadono nelle sezioni “dedicate” all’organizzazione della vacanza e quella della scelta della destinazione.

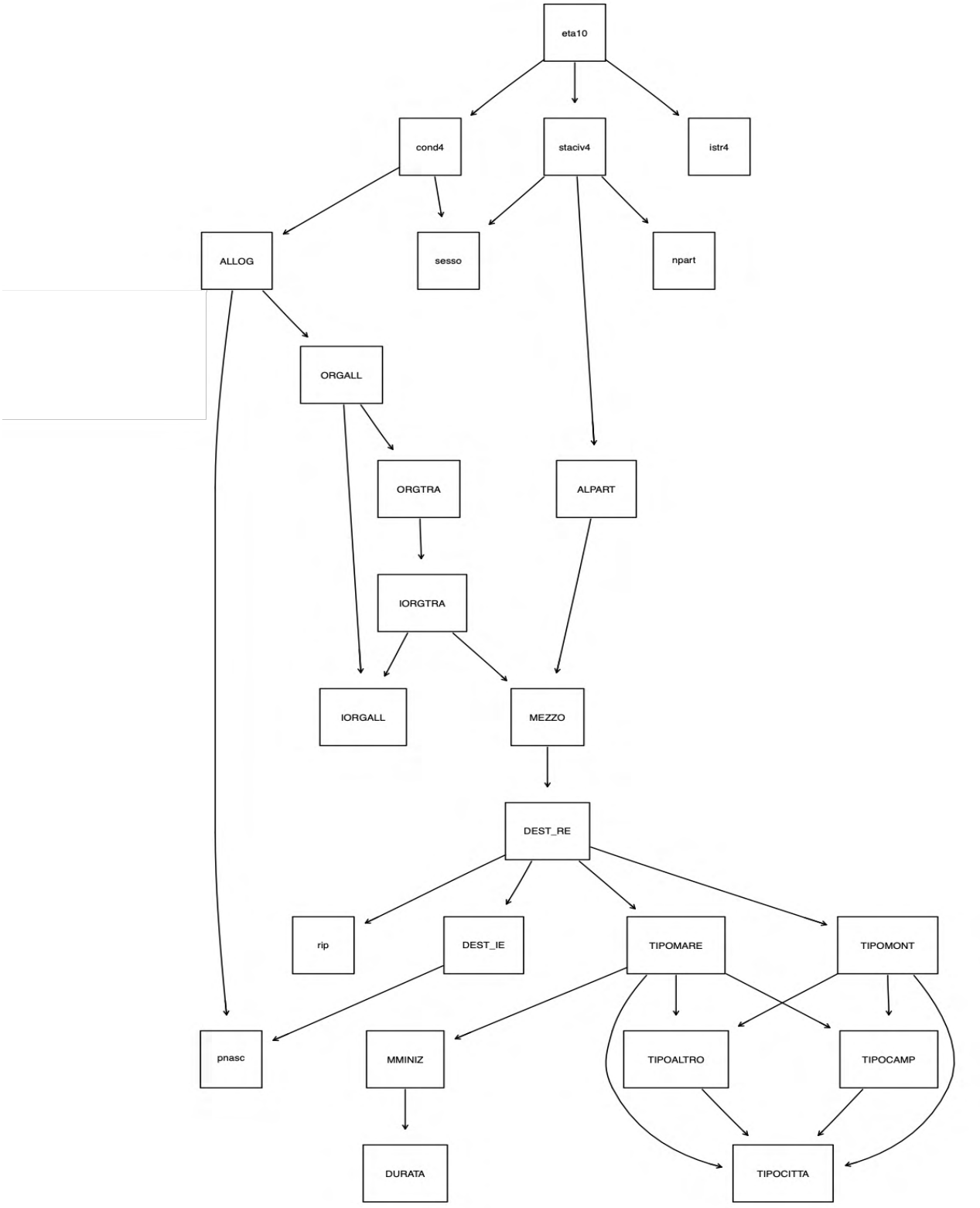


Figura 3.3: Rete Bayesiana relativa al dataset del 2014

Un confronto visivo rende chiaro quanto appena affermato; in un contesto puramente turistico, si conferma, inoltre, la ricorrenza nella scelta delle principali tipologie di destinazione, le quali ruotano sempre attorno a mare o montagna.
In ultimo di evidenzia, per il gruppo di variabili pertinenti all'organizzazione della vacanza, la presenza a distanza di molti anni del medesimo di archi seguenti la medesima direzione: dalla definizione dell'alloggio fino alla scelta del mezzo utilizzato.
La natura di questa similitudine è presumibilmente prodotta dal fatto che nel 2014 non si era ancora sviluppata una sorta di la fluidità nel settore turistico, la quale, grazie alla crescente popolarità e diffusione di servizi come AirB&B, ha reso sempre più facile l'incontro tra domanda e offerta di alloggia; di conseguenza nel 2014 si preferiva ancora viaggiare con più certezza nella definizione dell'alloggio, utilizzando, solo secondariamente, determinati canali organizzativi.

Riassumendo si può quindi dire che: il turismo è un fenomeno piuttosto stabile nelle sue dinamiche, le quali, tuttavia, possono mutare rapidamente in presenza di eventi esogeni, come la pandemia, ma che, nonostante ciò, non cambiano le relazioni di fondo.

3.6 Considerazioni sul settore del turismo

Attraverso tutti gli strumenti introdotti finora, è possibile formulare delle considerazioni sul fenomeno del turismo in senso stretto e apportare anche un confronto con le sue problematiche, delle quali si era brevemente accennato nei capitoli introduttivi.
Lo studio congiunto delle reti Bayesiane e delle statisrtiche descrittive ha confermato alcuni dei problemi emersi in questo settore. Innanzitutto, partendo dal problema più evidente connesso alla forte stagionalità del fenomeno, questa si presenta chiaramente evidenziata nell'analisi descrittiva del dataset: emerge, infatti, una notevole predilezione per le vacanze nei mesi estivi, che ,vista la correlazione con la stagione, vertono principalmente su mete marittime, di ciò si ricava ulteriore prova analizzando le reti Bayesiane, le quali denotano sempre una dipendenza diretta tra le variabili "tipo mare" e "mese di inizio viaggio".
Quantunque sia emerso che il fenomeno vacanziero graviti principalmente attorno ai mesi estivi, si evidenziano ulteriori stagionalità ricorrenti con il periodo natalizio e in

concomitanza con l'arrivo della primavera nel mese di aprile (si noti come nel corso del 2020 quest'ultimo slot sia stato completamente annullato dal lockdown imposto ai cittadini italiani).

Per quanto riguarda le destinazioni, invece, questo studio congiunto ha mostrato dei risultati un po' contrastanti. Sebbene la pandemia ha portato più eterogeneità tra le tipologie di destinazione, quelle marittime sono pur sempre "in testa", ma le alternative hanno acquistato più rappresentanza: una conferma di ciò è ottenibile confrontando la numerosità degli archi presenti nell'area dedicata alla tipologia della destinazione per il network del 2020, si evidenziano, infatti, molti archi diretti e pochi intrecci tra le varie variabili[24], le quali discendono più elegantemente dal nodo "tipo mare", raggruppandosi, peraltro, nel nodo "foglia" "tipo montagna", segno di una parziale inversione di tendenza in cui la vacanza in montagna si presentava sempre come principale alternativa a quella in località balneari.

Interessante notare come nonostante le mete di città siano scelte da una buona percentuale di intervistati, anche in misura superiore a quelle montane, appaiono nelle reti quasi sempre come un'alternativa di fondo. Una giustificazione di ciò può essere espressa in relazione alla durata, in quanto, ponendo come criterio per valutare il "peso" di una vacanza la durata della stessa, si è portati a pensare che la ricorrente posizione del nodo "tipo città" all'estremità dell'area topologia relativa alla tipologia della vacanza sia dovuta alla brevità di questa: difatti, i notevoli cali nei prezzi del trasporto e soluzioni sempre più veloci (vedasi la diffusione dell'alta velocità) hanno reso un'abitudine piuttosto generalizzata visitare determinate città durante i weekend.

Valutando queste preferenze nelle destinazioni ad un livello più dettagliato, analizzando la distribuzione regionale dei flussi è possibile delineare alcune considerazioni: non sempre una meta marittima aumenta le sue affluenze all'aumentare dell'apprezzamento della variabile "tipo mare", un esempio di ciò è il caso della Puglia che dal 2014 vede un calo di affluenze; inoltre, non è sempre vero che mete meno "turistizzate" siano in costante declino, regioni meno "popolari" come Abruzzo e Molise son state, infatti, in grado di aumentare le loro affluenze. Quindi, sebbene l'accentramento del turismo sia forte in determinate località (ad esempio Toscana, Trentino, Emilia-Romagna, Veneto, Lazio, Campania e isole) vi sono, comunque, delle aperture marginali di questo settore

[24] vedasi la Figura 3.3 per un confronto con una situazione più complessa

all'ampia platea di possibili offerte turistiche Italiane: fortunatamente non sempre vi è un disconoscimento delle bellezze turistiche Italiane meno in voga.

In relazione alle dinamiche di spesa, si era parlato di una race to the bottom per i prezzi delle offerte turistiche. Tuttavia, l'analisi pratica delle varie distribuzioni di densità di probabilità sulle variabili di spesa ha mostrato delle contraddizioni: ad eccezione del 2014 in cui tale variabile era assente, la valutazione delle dinamiche intertemporali per questa variabile mostra una inflazione abbastanza consistente per il settore. Questo scenario è probabilmente causato da un contesto di piena saturazione dell'offerta turistica, che combinata con l'incalzante eccesso di domanda, ha portato a degli aumenti per i prezzi di questo settore.

Ciò che non cambia con lo scenario esposto nell'introduzione concernente la tematica della sostenibilità: indipendentemente dall'andamento dei prezzi, un settore in "overbooking", fortemente stagionalizzato e ricorrente nelle mete, impoverisce i luoghi in cui si sviluppa, consumati dall'esigenza di sopperire e sopravvivere allo sfruttamento. Le ricadute sui prezzi sono solo l'ultima eventualità di questa spirale negativa, che, se combinata con l'inflazione, può contribuire a rendere questo settore sempre più "esclusivo" e meno accessibile alle varie stratificazioni sociali, potendo, quindi, scaturire un acuirsi di alcuni gap sociali.

In ogni caso, ragionando ora sull'utilità informativa portata da questa analisi nei soli limiti del contesto turistico, la rete Bayesiane, espletando il processo attuativo dell'agente vacanziero, danno la possibilità agli operatori del settore di conoscere l'andamento sequenziale di questi processi e, sulla base di questo, operare le corrette azioni di marketing o di posizionamento dell'offerta turistica.

Sapere che un turista definisce il mezzo prima dell'alloggio o conoscere quali variabili anagrafiche muovono più influenze sulla capacità di spesa, sono tutte informazioni che, razionalizzate dai network Bayesiani, permettono di ottenere maggiori vantaggi per gli operatori del settore.

Capitolo 4
Le reti Bayesiane come modello a supporto dell'analisi socioeconomica

Il passo successivo di questa analisi consiste: nel combinare una statistica descrittiva mirata con quanto emerso dal processo di elaborazione delle reti Bayesiane, per andare ad esaminare se il turismo presenta forme di stratificazioni sociali al suo interno.
Per valutare ciò, si farà riferimento a quanto è stato evidenziato dal "BES", il quale ha mostrato principalmente due assi di stratificazione sociale: quello di genere e quello geografico.
Sulla base di queste premesse, si andranno a identificare quattro individui tipo: uno di sesso maschile, uno di sesso femminile, uno residente in Nord Italia e uno residente nel Meridione (gli ultimi due individui non vengono suddivisi ulteriormente in base al sesso poiché l'informazione ricercata si basa solo sulle valutazioni di provenienza territoriale) per scoprire se quanto emerso dal BES si ritrova anche in questo settore o se vi siano particolari eccezioni.
In termini pratici per fare tutto ciò si opererà applicando ad ogni dataset le evidenze che permetteranno di individuare i quattro individui tipo, successivamente, mediante una statistica descrittiva sui nuovi dataset ottenuti, sarà possibile valutare, grazie a delle tabelle comparative, le distribuzioni relative per le variabili di interesse.
Il punto di partenza, perciò, per confrontare le informazioni di cui si è in possesso con le affermazioni del BES consiste nell'apportare i corretti filtri al dataset; l'applicazione di questi permette, infatti, di modellare le varie campionature per ottenerne delle altre in cui i dati presenti al loro interno riguardano esclusivamente l'individuo tipo che si vuole studiare.
Per essere più chiari, la restrizione imposta sulla variabile "sesso" permetterà di ottenere i due sub-dataset riguardanti gli individui di sesso femminile e quelli di sesso maschile. Allo stesso modo si otterranno quelli per la divisione Nord-Sud, solamente che la variabile d'interesse per discriminare i due sub-dataset sarà quella relativa alla ripartizione geografica dei soggetti intervistati.
Una volta fatto ciò e individuate correttamente le quattro "sotto campionature", si procede con l'esplicitazione della statistica descrittiva per valutare e confrontare gli andamenti di alcune variabili chiave.

In questa comparazione non saranno esaminate tutte le variabili presenti nei vari records ma sarà effettuata una sorta di cherry picking tra esse, al fine di valutare solo quelle che si ritiene essere più pertinenti alla sfera sociale.

Sulla base di questa premessa sono state quindi individuate come più rappresentative le variabili: istruzione, condizione professionale soggettiva, luogo di destinazione vacanza, destinazione Italia o Estero (sebbene queste ultime due variabili rispetto alle prime hanno una carica informativa poco pertinente con la sfera sociale, la loro presenza è propedeutica alla contestualizzazione "turistica" di quest'analisi, nonché può essere di supporto giustificativo per alcune considerazioni finali), numero di membri della famiglia partecipanti alla vacanza, quella indicante se un soggetto ha compiuto più viaggi durante l'anno e, infine, le variabili di spesa media contestualizzate con la durata della vacanza.

Stabilito quali variabili esaminare, l'analisi porrà a diretto confronto quanto rilevato per ognuno dei quattro soggetti archetipo individuati, in modo da rendere più chiare le eventuali disparità; inoltre, la possibilità di valutare queste variabili a distanza di anni, permetterà di stabile se eventuali trend o evidenze fanno parte di un processo più ampio.

L'analisi inizia, dunque, confrontando quanto emerge dalla variabile "istruzione": la tabella c. (sebbene vi siano le frecce indicanti il verso dell'andamento, rispetto alle tabelle precedentemente esaminate non si apporrà nessun giudizio a questo tramite il colore utilizzato, bensì, il colore sarà utilizzato esclusivamente per evidenziare quale delle categorie presenta le peggiori o migliori rilevazioni oppure se si vuole mettere in risalto determinati dati), qui riportata, mostra come si è evoluta l'istruzione dei soggetti intervistati.

Limitatamente al primo scaglione, le evidenze confermano quanto espresso a livello generale nel BES, ovvero un tendenziale primato del Nord e il sesso femminile non al passo con la controparte.

Questo risulta vero se ci si limitasse a valutare l'anno 2014, tuttavia, l'analisi dinamica permette di notare come lo scenario iniziale nel corso del tempo si sia completamente ribaltato: nel 2020, invero, donne e residenti nel Sud Italia mostrano un'inversione di tendenza, ottenendo i risultati migliori per le rispettive categorie[25].

Per gli scaglioni centrali non si esprimono particolari considerazioni, in quanto le fluttuazioni si compensano vicendevolmente e i vari equilibri restano costanti negli anni;

[25] Si inserisce un link al sito ufficiale delle Prove Invalsi per ulteriori approfondimenti al contesto di riferimento sull'istruzione: https://www.invalsiopen.it/poverta-educativa-questione-meridionale/

unica nota a riguardo sta alla presenza per l'anno 2020 di un ottimo livello di istruzione secondaria superiore per la categoria Sud, nota di positiva controtendenza rispetto agli anni precedenti.

In ultimo, per lo scaglione pertinente all'istruzione più elevata, emergono dei dati apparentemente in contrasto con quanto ci si sarebbe potuto attendere: donne e cittadini del Sud si presentano per tutti gli anni come più istruiti delle controparti; dunque, non è sempre vero che queste categorie risultano penalizzate; tuttavia, bisogna stabilire se gli elevati livelli di istruzione comportino anche benefici in altri contesti[26].

c.	Istruzione							
		2014		2018		2019		2020
nulla/elem	Uomini:	4,72%	↘	3,20%	↘	2,06%	↗	2,08%
	Donne:	7,80%	↘	3,45%	↘	3,07%	↘	2,07%
	Nord:	6,43%	↘	3,42%	↘	2,90%	↘	2,01%
	Sud:	6,61%	↘	3,50%	↘	1,79%	↗	1,80%
media/avv. Prof.	Uomini:	22,83%	↗	23,22%	↗	23,48%	↘	20,93%
	Donne:	20,88%	↘	17,46%	↗	18,91%	↘	18,17%
	Nord:	24,08%	↘	21,42%	↗	22,37%	↘	20,68%
	Sud:	19,63%	↘	18,66%	↗	19,64%	↘	16,67%
second.sup o qual prof.	Uomini:	44,49%	↗	46,44%	↗	46,97%	↗	47,12%
	Donne:	42,55%	↗	46,50%	↘	44,34%	↗	46,14%
	Nord:	43,86%	↗	47,17%	↘	46,74%	↗	48,58%
	Sud:	42,56%	↗	44,02%	↗	46,07%	↗	49,10%
diplome univ. Post laurea	Uomini:	27,95%	↘	27,14%	↗	27,49%	↗	29,87%
	Donne:	28,77%	↗	32,59%	↗	33,69%	↘	33,62%
	Nord:	25,63%	↗	28,00%	↘	28,00%	↗	28,73%
	Sud:	31,20%	↗	33,82%	↘	32,50%	↘	32,43%

Tabella c: La tabella presenta le distribuzioni relative ai vari livelli di istruzione per i quattro individui tipo, il colore rosso è utilizzato per evidenziare quale delle categorie presenta le peggiori rilevazioni, il verde per indicare le migliori, mentre l'azzurro per mettere in risalto determinati dati.

In termini di valutazione delle condizioni professionali, la tabella d. restituisce uno scenario in cui le divergenze tra i quattro individui tipo appaiono più marcate; in termini occupazionali, donne e cittadini del Sud appaiono notevolmente penalizzati.

Fortunatamente dall'anno 2014 sono stati ottenuti dei miglioramenti che hanno contribuito ad appianare queste divergenze: questo lo si nota molto per i residenti del meridione, che in pochi anni hanno quasi colmato il gap con quelli del Nord*, attestandosi su percentuali pressoché analoghe; anche le donne mostrano andamenti positivi, anche se

[26] Si rimanda ad un articolo sulle disuguaglianze di genere: https://www.openpolis.it/le-troppe-disuguaglianze-di-genere-nellistruzione/

mantengono alte percentuali nella categoria relativa ai soggetti non partecipanti al mercato del lavoro (in quanto casalinghe, studentesse o altro).

Per la categoria pensionati o ritirati dal lavoro, si notano percentuali molto basse per il Sud se paragonate ai dati "settentrionali". Il motivo di ciò non è facilmente identificabile, ci si limita a suggerire come cause: o la presenza di un mercato lavorativo ancora "giovane", dati solo i recenti aumenti dell'occupazione, o la necessità per soggetti in età avanza di mantenere un impiego per avere gli adeguati mezzi di sostentamento; si sottolinea per il Sud, anche, una percentuale abbastanza elevata e costante per la categoria in cerca di lavoro, segno di una significativa difficoltà di accesso a tale mercato.

Per quanto riguarda l'impatto pandemico, infine, le evidenze mostrano forti ricadute per le fasce "deboli" (Sud e donne), le quali perdono parte progressi ottenuti nel 2019, nello specifico: il Sud rileva evidenti cali nell'occupazione mentre le donne, pur mantenendo stabile il trend di crescita occupazionale, segnalano aumenti per la categoria dei soggetti non partecipanti al mercato del lavoro.

d.	condizione professionale soggettiva							
		2014		2018		2019		2020
occupato	Uomini:	60,14%	↗	63,26%	↗	63,74%	↗	66,22%
	Donne:	47,75%	↗	56,68%	↘	54,22%	↗	57,13%
	Nord:	58,50%	↗	60,75%	↘	59,29%	↗	63,86%
	Sud:	40,08%	↗	54,23%	↗	57,14%	↘	50,45%
in cerca lav.	Uomini:	5,02%	↘	3,72%	↗	4,11%	↘	3,79%
	Donne:	8,23%	↘	5,51%	↘	5,47%	↘	5,33%
	Nord:	4,96%	↘	3,33%	↗	3,70%	↘	3,11%
	Sud:	11,16%	↘	10,20%	↘	10,00%	↗	11,71%
casalinga/stud/altro	Uomini:	10,63%	↗	11,76%	↘	11,04%	↘	9,30%
	Donne:	28,16%	↘	21,48%	↗	24,18%	↘	20,78%
	Nord:	16,44%	↘	14,83%	↗	17,86%	↘	13,17%
	Sud:	29,55%	↘	21,57%	↗	21,79%	↗	25,23%
pensionato/ritirato	Uomini:	24,21%	↘	21,26%	↘	21,10%	↘	20,69%
	Donne:	15,86%	↗	16,34%	↘	16,12%	↗	16,76%
	Nord:	20,10%	↗	21,08%	↘	19,15%	↗	19,85%
	Sud:	19,21%	↘	13,99%	↘	11,07%	↗	12,61%

Tabella d: Composizione delle condizioni lavorative dei quattro individui tipo, il colore rosso è utilizzato per evidenziare quale delle categorie presenta le peggiori rilevazioni, il verde per indicare le migliori, mentre l'azzurro per mettere in risalto determinati dati.

Spostando, ora, l'analisi su un piano più pertinente con il turismo si andranno ad analizzare i principali luoghi di vacanza per questi individui tipo; in questo contesto perde di capacità informativa il dualismo uomo-donna, tuttavia, le differenze locative dei soggetti intervistati esemplificano, nella tabella e., alcune differenziazioni di settore.

Dalle analisi effettuate nei capitolo precedente si nota una forte dipendenza tra tipologia delle destinazioni e ripartizione geografica del soggetto[27]: sebbene il Nord Italia presenti un offerta di destinazioni più variegata rispetto al Sud (il quale, per conformazione geografica, si caratterizza ad esempio per la concentrazione di numerose località marittime a scapito di quelle montane) si può notare per i cittadini del Meridione una notevole predilezione per le località di mare e cittadine più tipiche delle loro zone (tabella e); questo fenomeno di predilezione delle località tipiche del proprio, risulta, probabilmente, presente anche nel Nord, però, è più difficile da valutare per via dell'offerta variegata di possibili destinazioni "tipo" che presenta l'Italia Settentrionale.
Tutto ciò evidenzia un turismo molto eradicato nel territorio limitrofo agli intervistati. La minor eterogenia nelle destinazioni può essere, per contro, considerata come una sorta di termometro delle capacità di spesa: presumendo la distanza del viaggio come variabile che ne aumenta i costi, la fruizione di mete vicine potrebbe implicare un contenimento delle spese o una minor capacità economica per il soggetto intervistato.
Se ciò risulta trasversalmente vero per ogni tipologia di individuo, quanto detto accresce di rilevanza, quando in seguito sarà confrontato con altre valutazioni, più correlate ad un contesto economico puro e in grado di far emergere alcune differenze sostanziali.

[27] Si rimanda ad un confronto con le reti Bayesiane presentate nel capitolo 3 e in appendice per confermare la premessa, in quanto la presenza di archi diretti tra la variabile "rip" e "DEST_RE", giustifica pienamente queste affermazioni

e.	Luogo di destinazione vacanza							
		2014		2018		2019		2020
mare	Uomini:	30,76%	↗	35,39%	↗	37,59%	↗	39,22%
	Donne:	29,76%	↗	36,48%	↗	36,94%	↗	39,77%
	Nord:	30,48%	↗	36,06%	↗	37,87%	↘	36,38%
	Sud:	32,45%	↗	35,86%	↗	38,92%	↗	50,33%
montagna	Uomini:	17,50%	↗	20,37%	↘	17,79%	↗	23,66%
	Donne:	16,93%	↗	18,20%	↘	18,03%	↗	22,47%
	Nord:	22,14%	↗	23,34%	↘	20,59%	↗	27,30%
	Sud:	7,27%	↗	9,20%	↘	8,38%	↗	11,92%
città	Uomini:	38,04%	↘	29,79%	↗	33,17%	↘	23,11%
	Donne:	40,09%	↘	31,05%	↗	33,19%	↘	24,41%
	Nord:	34,58%	↘	26,43%	↗	30,02%	↘	21,34%
	Sud:	47,34%	↘	41,15%	↘	37,84%	↘	30,46%
campagna	Uomini:	9,71%	↗	9,74%	↘	8,27%	↗	11,92%
	Donne:	8,81%	↗	10,40%	↘	8,76%	↗	11,48%
	Nord:	10,06%	↗	10,35%	↘	8,82%	↗	13,01%
	Sud:	6,21%	↗	8,51%	↘	8,11%	↘	6,62%
altro	Uomini:	3,99%	↗	4,71%	↘	3,17%	↘	2,09%
	Donne:	4,40%	↘	3,86%	↘	3,09%	↘	1,86%
	Nord:	2,73%	↗	3,82%	↘	2,70%	↘	1,96%
	Sud:	6,74%	↘	5,29%	↗	6,76%	↘	0,66%

Tabella e: La tabella mostra le preferenze dei quattro soggetti esaminati in relazione ai tipo di destinazione presenti nei dataset.

Ulteriore profondità al ragionamento che collega la vicinanza della destinazione con le capacità di spesa[28], si ottiene dalla tabella f., la quale ci permette di capire la percentuale di individui che si spostano all'estero per effettuare una vacanza.

Sebbene uscire dai confini italiani sia oggettivamente più semplice per i cittadini del Nord, la premessa che spostarsi all'estero, cioè allontanarsi dai propri luoghi limitrofi, comporti una vacanza più costosa, mostra le prime divergenze di sfumatura economica tra i vari individui tipo.

Si può notare, infatti, un divario abbastanza sostanziale tra Nord e Sud, il quale nel tempo è andato, tuttavia, per ridursi; confermando, quindi, la generica visione di un Sud economicamente meno solido.

[28] Questa relazione è visivamente apprezzabile nelle reti Bayesiane create, dove il gruppo di nodi pertinenti alla tipologia della destinazione ha sempre ricadute dirette sulle variabili di spesa

f.	destinazione Italia o estero							
		2014		2018		2019		2020
Italia	Uomini:	89,17%	↗	89,58%	↘	85,39%	↗	94,74%
	Donne:	87,95%	↗	88,80%	↘	84,55%	↗	95,87%
	Nord:	85,35%	↗	87,50%	↘	83,35%	↗	94,42%
	Sud:	92,15%	↗	93,88%	↘	90,00%	↗	95,95%
Estero	Uomini:	10,83%	↘	10,42%	↗	14,61%	↘	5,26%
	Donne:	12,05%	↘	11,20%	↗	15,45%	↘	4,13%
	Nord:	14,65%	↘	12,50%	↗	16,65%	↘	5,58%
	Sud:	7,85%	↘	6,12%	↗	10,00%	↘	4,05%

Tabella f: La tabella mostra la frequenza percentuale di viaggia all'Estero o in Italia per i quattro individui tipo.

Un ulteriore criterio per valutare le capacità economiche del turista si basa sul capire quanti soggetti partecipano con lui alla vacanza (tabella g.); tale premessa valutativa è ottenuta sia da principi di ragionamento logico, sia, soprattutto, dalle informazioni espresse dalle reti Bayesiane, le quali vedono legami diretti tra le variabili di spesa e il nodo "npart", che identifica, per l'appunto, il numero di famigliari partecipanti alla vacanza.[29]

Come si può notare, limitatamente al numero di famigliari coinvolti, non appaiono particolari evidenze, la situazione risulta pressoché equamente bilanciata tra Nord e Sud (come anche tra uomo e donna). Per entrambi i casi, il nucleo famigliare partecipante al viaggio si compone, nella maggior parte delle rilevazioni, di almeno tre membri dell'apparato parentale; non emergono, perciò, particolari indicazioni da differenziare fortemente alcuno dei quattro individui tipo esaminati.

[29] Si rimanda ad un confronto con le reti Bayesiane presentate nel capitolo 3 e in appendice

g.		numero di membri della famiglia partecipanti alla vacanza							
		2014		2018		2019		2020	
1	Uomini:	34,06%	↘	27,24%	↗	28,79%	↘	25,09%	
	Donne:	34,23%	↘	28,20%	↗	29,65%	↘	25,46%	
	Nord:	32,71%	↘	27,25%	↗	29,12%	↘	25,80%	
	Sud:	40,29%	↘	30,32%	↗	31,07%	↘	24,32%	
2	Uomini:	36,81%	↗	39,42%	↗	41,34%	↘	39,66%	
	Donne:	37,61%	↗	38,47%	↗	39,92%	↘	39,50%	
	Nord:	38,08%	↗	40,75%	↘	39,58%	↘	39,34%	
	Sud:	33,68%	↗	35,28%	↗	38,93%	↘	36,94%	
3	Uomini:	14,96%	↗	16,31%	↘	13,85%	↗	17,50%	
	Donne:	14,99%	↗	16,43%	↘	15,45%	↗	19,48%	
	Nord:	16,11%	↗	17,08%	↘	15,69%	↗	16,74%	
	Sud:	14,67%	↘	8,75%	↗	12,14%	↗	20,72%	
4	Uomini:	12,01%	↗	14,65%	↘	13,64%	↗	14,81%	
	Donne:	11,53%	↗	14,47%	↘	13,05%	↗	13,38%	
	Nord:	10,98%	↗	12,75%	↗	13,60%	↗	15,37%	
	Sud:	9,92%	↗	18,95%	↘	16,43%	↗	17,12%	
5	Uomini:	2,17%	↗	2,37%	↗	2,38%	↗	2,94%	
	Donne:	1,65%	↗	2,43%	↘	1,92%	↗	2,18%	
	Nord:	2,12%	↗	2,17%	↘	2,01%	↗	2,74%	
	Sud:	1,45%	↗	6,71%	↘	1,43%	↘	0,90%	

Tabella g: La tabella mostra, in base a degli scaglioni, le frequenze relative del numero di soggetti partecipanti alla vacanza per i quattro individui tipo.

Ultimo criterio per valutare le capacità economiche, prima di passare alle più dirette informazioni sulla spesa, riguarda il numero di viaggi svolti dall'individuo tipo durante l'anno delle rilevazioni: anche qui la differenziazione uomo donna non è di gran valore, poiché le loro percentuali sono pressoché sovrapponibili; perciò, si farà più attenzione a quella tra Nord e Sud.

Dalle osservazioni presenti nella tabella h., i cittadini residenti in Nord Italia si dimostrano più propensi a svolgere plurime vacanze durante l'anno rispetto alla controparte meridionale; diversamente da alcuni trend precedentemente discussi, l'analisi intertemporale di queste rilevazioni porta ad esprimere un giudizio negativo sull'andamento di queste variabili.

Dal 2014, infatti, probabilmente per un aumento dei costi necessari per effettuare una vacanza, si è significativamente ridotta la percentuale di soggetti pluri-vacanzieri; questo fenomeno ha colpito più il Nord che il Sud, segno, comunque, di un effetto di margine, per cui si è stati portati a fare economia risparmiando sulle vacanze "accessorie" piuttosto che su quella "principale" tipica dei mesi estivi[30].

[30] Si rivedano le considerazioni emerse nel capitolo 3 riguardanti la stagionalità del fenomeno turistico

A sostegno di questa argomentazione vi è l'analisi del singolo anno 2020 che, nonostante lo strascico pandemico, data la non saturazione dell'offerta turistica e il notevole calo di domanda, ha rilevato un aumento, tanto al Nord quanto al Sud, di soggetti pluri-vacanzieri.

h.	Più viaggi nel corso dell'anno							
		2014		2018		2019		2020
SI	Uomini:	7,69%	↘	6,65%	↘	3,83%	↗	5,41%
	Donne:	7,94%	↘	6,60%	↘	4,01%	↗	5,56%
	Nord:	8,94%	↘	5,87%	↘	3,93%	↗	5,02%
	Sud:	3,19%	↗	4,57%	↘	2,96%	↗	4,27%
NO	Uomini:	92,31%	↗	93,35%	↗	96,17%	↘	94,59%
	Donne:	92,06%	↗	93,40%	↗	95,99%	↘	94,44%
	Nord:	91,06%	↗	94,13%	↗	96,07%	↘	94,98%
	Sud:	96,81%	↘	95,43%	↗	97,04%	↘	95,73%

Tabella f: La tabella mostra la percentuale relativa degli individui tipo che hanno effettuato più viaggi nel corso di un anno.

Venendo ora a valutare dati più prettamente economici la tabella i. mette in evidenza le spese medie, giornaliere e complessive, dei quattro individui tipo, riportando, anche, un'indicazione della durata media delle vacanze per aumentare di significatività i dati sulla spesa complessiva, in quanto fortemente correlata alla durata della vacanza[31].
Valutando il binomio uomo donna, le rilevazioni fanno intuire andamenti pressoché equiparabili tra i vari pesi percentuali, di contesto si rilevano solo minime variazioni di spesa in favore degli uomini, mentre il 2019 si presenta, invece, come caso a sé, giustificabile dall'estemporanea durata maggiore per le vacanze sostenute dal sesso femminile; in ultimo vi è una nota negativa sull'aumento nel 2020 del gap di spesa tra i due generi, probabile evidenza di come la pandemia abbia danneggiato maggiormente le fasce "più delicate" della popolazione.
Nei confronti delle fattispecie Nord Sud, emergono, bene o male, le stesse considerazioni fatte per il binomio precedente; anche qui, il Nord si presenta come categoria "dominante" sia per capacità di spesa sia per la possibilità di dedicare più tempo alle vacanze.
Da un punto di vista dinamico, il Sud nel 2018 si presentava notevolmente distante dai valori ottenuti dalla controparte. Ad ogni modo, queste differenze sono state per la maggior parte colmate nel 2019, anno in cui, sia per durata che per spesa, i valori mostrati

[31] Si rimanda per la valutazione di tale evidenza agli elaborati Bayesiani, dove è possibile vedere, per tutti gli anni in esame, una connessione diretta tra le variabili "DURATA" ed "ESPE_CO"

tendono a convergere; per quanto riguarda il 2020, come avvenuto anche in altri contesti precedentemente analizzati, le ricadute della pandemia hanno parzialmente annullato quanto di ottenuto nel corso dell'anno precedente.

Confrontando quanto emerso da quest'ultima tabella (tabella i.) e le precedenti riguardanti l'istruzione e la condizione professionale (tabelle c. e d.) è possibile affermare che emergono correlazioni tra le possibilità di spesa e la condizione lavorativa[32]: per cui le fattispecie più "performanti" risultano per essere anche quelle più "economicamente abbienti"; tuttavia, se a questa valutazione si aggiungono le informazioni ricavate sull'istruzione, si denota la presenza di una sorta di "comportamento" discriminatorio per le categorie, impropriamente dette, "deboli"[33].
In altri modi quello che si vuole dire è che: non vi è premio, in termini di capacità di spesa nel contesto vacanziero e di rappresentanza occupazionale nel mercato del lavoro, per la maggior istruzione rilevata tra le donne e i cittadini residenti in Sud Italia.
La tabella mostra, inoltre, evidenze con le precedenti presupposizioni sull'inflazione di questo settore, sottolineando un significativo aumento dei prezzi culminato nel 2019, che ha causato: sia la predilezione di mete all'interno dei confini italiani, a conferma che le destinazioni limitrofe alla propria residenza rispondono a criteri di accessibilità economica, sia una riduzione delle percentuali di soggetti pluri-vacanzieri; inoltre, il contesto post-pandemico, seppur caratterizzato da prezzi piuttosto alti, non più attanagliato dalle precedenti pressioni sul settore turistico, vede, per le vacanze, un miglioramento del rapporto spesa-durata, a conferma della momentanea accessibilità al settore.

[32] Si inserisce l'URL che riporta ad un report dell'Istat in merito allo studio delle spese per i consumi delle famiglie italiane: https://www.istat.it/it/files/2021/06/REPORT_CONSUMI_FAMIGLIE_2020.pdf
[33] Si inserisce l'URL che riporta ad un report dell'Istat in merito allo studio della correlazione tra livelli di istruzione e ritorni occupazionali: https://www.istat.it/it/files/2020/07/Livelli-di-istruzione-e-ritorni-occupazionali.pdf

i.	spesa media sostenuta per la vacanza					
		2018		2019		2020
Complessiva	Uomini:	355,25	↗	384,09	↗	388,69
	Donne:	351,45	↗	404,15	↘	373,86
	Nord:	374,16	↗	401,88	↘	389,16
	Sud:	305,81	↗	395,35	↘	342,58
Giornaliera	Uomini:	90,00	↗	97,49	↘	86,29
	Donne:	88,10	↗	95,04	↘	84,01
	Nord:	90,82	↗	94,28	↘	87,86
	Sud:	86,08	↗	94,64	↘	70,79
durata media vac. In gg	Uomini:	4,92	↗	4,96	↗	5,59
	Donne:	4,99	↗	5,35	↗	5,56
	Nord:	5,08	↗	5,43	↗	5,60
	Sud:	4,58	↗	5,20	↗	5,72

Tabella i: La tabella evidenzia le spese medie complessive e giornaliere per i quattro individui tipo, indicando anche la durata media delle vacanze in giorni per questi soggetti.

In conclusione, ricordando che queste osservazioni sono comunque pertinenti ad un contesto turistico non accessibile ad ogni anfratto della società, l'analisi qui svolta mostra uno scenario dai tratti omogenei; le differenze emerse, seppur con il passare del tempo siano diventate più lievi, sviluppano una sorta di filo conduttore attraverso le variabili esaminate, presentandosi con ricorrenza in ogni tabella esaminata e delineando la conformazione di una flebile forma di stratificazione sociale, acuitasi in seguito alla pandemia.

Appare di oggettiva valutazione una sottile discriminazione per i soggetti appartenenti alle categorie: donne e residenti nel Sud; tale discriminazione si è mostrata sempre più velata con il passare degli anni ma alcuni suoi tratti appaiono ancora evidenti, ciò lo si vede, ad esempio: nei confronti dell'istruzione, che, come è stato illustrato in precedenza, non riconosce i meriti a chi si dimostra più preparato.

Inoltre, gli eventi accaduti durante il 2020, sebbene abbiano colpito e riguardato tutti, i loro effetti hanno avuto ricadute più forti per queste due categorie più "deboli", segno che, nonostante gli sviluppi positivi degli anni precedenti, queste differenze sono più profondamente radicate di quanto si pensi, poiché, in un contesto in cui i principi di equità sociale si presentano come consolidati, le ricadute della pandemia sarebbero dovute essere distribuite tra le varie categorie di soggetti analizzati in maniera equa e non asimmetricamente come è avvenuto all'atto pratico.

Dal punto di vista della pura fruizione del proprio tempo libero, nonostante le discrepanze di spesa e nelle destinazioni, la vacanza si presenta come fenomeno sociale proprio della cultura moderna, che si adatta alle possibilità dell'individuo e non viene meno neppure in contesti di sofferenza sociale ed economica; in un certo senso, quindi, è una sorta di necessità che va oltre discriminazioni geografiche o di genere.

CAPITOLO 5
Conclusioni

In conclusione, l'utilizzo delle reti Bayesiane per descrivere quanto ci fosse di celato dietro a questi dataset, ha confermato le aspettative sulle proprietà del modello; questi network si dimostrano ancora una volta capaci di restituire un'informazione chiara rispetto a problemi complessi, le diramazioni ottenute sulla base di relazioni di probabilità condizionata, come è stato mostrato, sono state propedeutiche alla formulazione di un pensiero critico per le analisi svolte.

Il confronto tra le statistiche descrittive e le reti create ha permesso di individuare correlazioni chiave all'interno di questo contesto turistico, le quali sono state, successivamente, utilizzate a supporto dell'analisi socioeconomica.

Tutto ciò è stato reso concepibile, anche, grazie all'immediatezza con cui queste reti sono in grado di trasmettere le informazioni che racchiudevano le campionature esaminate, offrendo un punto di vista differente sul problema.

La convergenza tra le evidenze ottenute con gli esiti di altri studi di diversa provenienza, come il BES e quelli elencanti le problematiche del turismo, è di ulteriore testimonianza alla bontà analitica di questo approccio scientifico.

Un raffronto tra i dati a disposizione e i network costruiti ha reso inoltre possibile, per quanto poco, aumentare la consapevolezza sui problemi studiati, evidenziando qualche sfumatura che altrimenti non sarebbe stata colta.

"Possiamo lamentarci perché i roseti hanno le spine o rallegrarci perché i cespugli spinosi hanno le rose. Dipende dai punti di vista".
(Abraham Lincoln)

Bibliografia & Sitografia

Boella M., (2011), *Probabilità e Statistica per ingegneria e scienze*, Parson Italia s.p.a., Milano

Korb & Nicholson (2004), *Bayesian Artificial Intelligence*, Chapman & Hall/CRC, Boca Raton, Florida

Kjaerulff & Madsen (2008), *Bayesian Network and Influence Diagrams A Guide to Construction and Analysis*, Springer Science+Business Media, New York

Scutari & Denis (2015), *Bayesian Networks with examples in R*, Taylor & Francis Group, New York

Nagarajan R., Scutari M. & Lèbre S. (2013), *Bayesian Network in R with Application in Systems Biology*, Springer Science+Business Media, New York

Pourret O., Naim P. & Marcot B. (2008), *Bayesian Networks A Practical Guide to Application*, John Wiley & Sons, Ltd, Southern Gate, West Sussex, England

De Jonge E. & Van der Loo M. (2013), *An introduction to data cleaning with R*, Statistics Netherlands, The Hague/Heerlen

Cooper G. & Herskovits E. (1992), *Bayesian Method for the Induction of Probabilistic Networks from Data*, Kluwer Academic Publishers, Boston

De Campos L. M. (2006), *A scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Test*, Nir Friedman

Istat (2020), *BES 2020 il benessere equo e sostenibile in Italia*, disponibile a https://www.istat.it/it/files//2021/03/BES_2020.pdf

Istat (2021), *BES dieci anni di misurazione del benessere equo e sostenibile*, disponibile a https://www.istat.it/it/files//2021/03/BES_2020-nota-stampa.pdf

Istat (2020), *Livelli di istruzione e ritorni occupazionali anno 2019*, disponibile a https://www.istat.it/it/files/2020/07/Livelli-di-istruzione-e-ritorni-occupazionali.pdf

Istat (2021), *Le spese per i consumi delle famiglie anno 2020*, disponibile a https://www.istat.it/it/files/2021/06/REPORT_CONSUMI_FAMIGLIE_2020.pdf

Istat (2021), *Indagine viaggi e vacanze Aspetti metodologici dell'indagine anno 2020*, disponibile a https://www.istat.it/microdata/download.php?id=/import/fs/pub/wwwarmida/264/2020/01/Nota.pdf

INVALSIopen (2020), *Povertà educativa: esiste una questione Meridionale?*, disponibile a https://www.invalsiopen.it/poverta-educativa-questione-meridionale/

Openpolis (2019), *Troppe disuguaglianze di genere nell'istruzione*, disponibile a https://www.openpolis.it/le-troppe-disuguaglianze-di-genere-nellistruzione/

Ministero dei beni e delle attività culturali e del turismo, *La strategia italiana per il turismo sostenibile*, disponibile a https://www.beniculturali.it/mibac/multimedia/MiBAC/documents/1443695985552_3-_La_strategia.pdf

Enit (2020), *Rapporto sui Risultati 2019*, disponibile a https://www.enit.it/wwwenit/images/amministrazionetrasparenteepe/Bilancio2020/consuntivo/7.%20Rapporto%20Risultati%20da%20PIRAB%202019-1.pdf

FormazioneTurismo.com, *La Strategia dell'Italia per il Turismo Sostenibile e per lo Sviluppo del Sud*, disponibile a https://academy.formazioneturismo.com/la-strategia-dellitalia-per-il-turismo-sostenibile-e-per-lo-sviluppo-del-sud/

Econopoly (2019), *Presente e futuro del turismo, tutte le tendenze. E la lentezza dell'Italia*, Il Sole 24ore, 15/05/2019

Istat (2021), Microdati, Viaggi e vacanze: file ad uso pubblico, disponibili a https://www.istat.it/it/archivio/178695

bnlearn, https://www.bnlearn.com/

APPENDICE

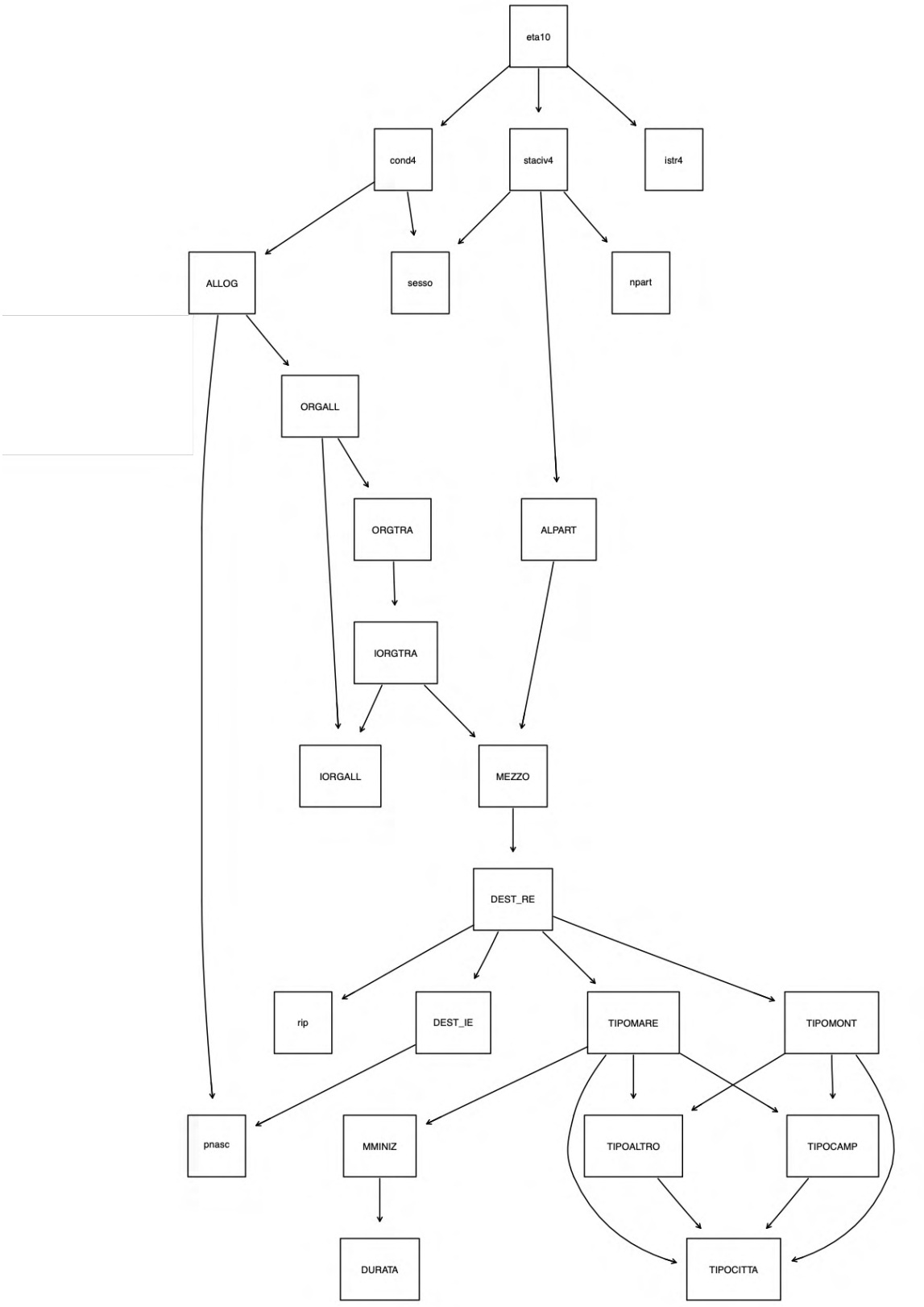
Elenco delle variabili presenti nei dataset:

Descrizione variabile:	**Acronimo:**	**Modalità:**
Sesso del soggetto intervistato	sesso	2
Classe di età del soggetto intervistato	eta10	8
Stato civile del soggetto intervistato	staciv4	4
Paese di nascita del soggetto intervistato	pnasc	2
Ripartizione geografica di residenza del soggetto in intervistato	rip	5
Titolo di studio più altro conseguito dal soggetto intervistato	istr4	4
Condizione professionale soggettiva dell'intervistato	cond4	4
Regione italiana o stato europeo o macroarea geografica non UE di destinazione principale del viaggio	DEST_RE	51
Destinazione principale Italia o Estero	DEST_IE	2
Mese di inizio del viaggio	MMINIZ	12
Durata del viaggio in numero di notti	DURATA	Numero Cardinale
Principale mezzo di trasporto utilizzato dal soggetto intervistato	MEZZO	11

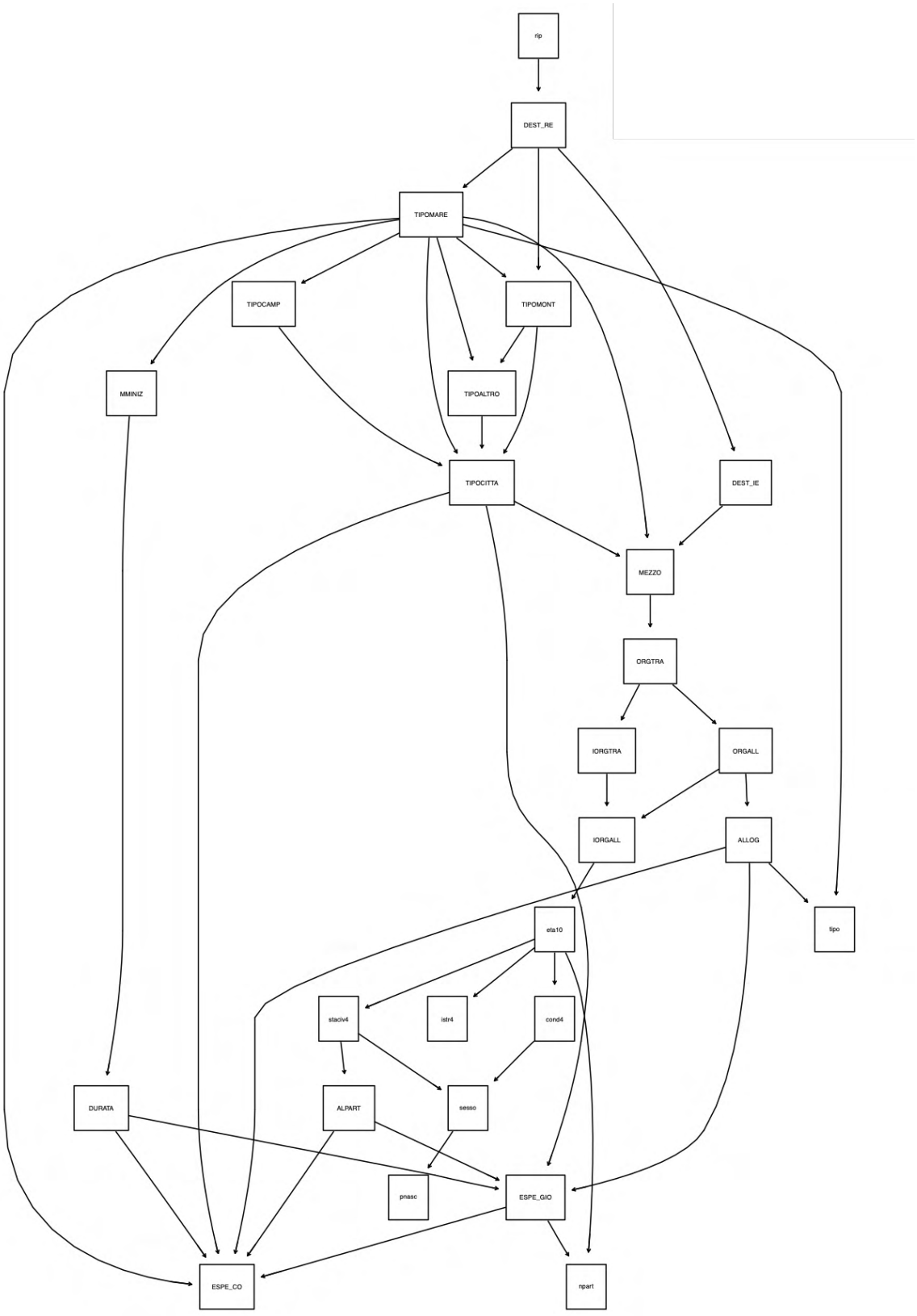
Principale tipo di alloggio utilizzato dal soggetto intervistato	ALLOG	18
Organizzazione dell'alloggio	ORGALL	4
Organizzazione del trasporto	ORGTRA	4
Filtro: se hanno partecipato alla vacanza altri componenti della famiglia dell'intervistato	ALPART	2
Utilizzo di Internet per prenotare l'alloggio	IORGALL	3
Utilizzo di Internet per prenotare il trasporto	IORGTRA	3
Luogo tipo di destinazione della vacanza (mare)	TIPOMARE	2
Luogo tipo di destinazione della vacanza (montagna)	TIPOMONT	2
Luogo tipo di destinazione della vacanza (città)	TIPOCITTA	2
Luogo tipo di destinazione della vacanza (campagna)	TIPOCAMP	2
Luogo tipo di destinazione della vacanza (altro)	TIPOALTRO	2
Numero di componenti della famiglia partecipanti al viaggio espresso in classi	npart	5
Tipologia di attività principale svolta nel viaggio	tipo	8
Spesa media sostenuta per l'intero viaggio	ESPE_CO	Indicazione numerico-quantitativa

Spesa media giornaliera sostenuta durante il viaggio	ESPE_GIO	Indicazione numerico-quantitativa

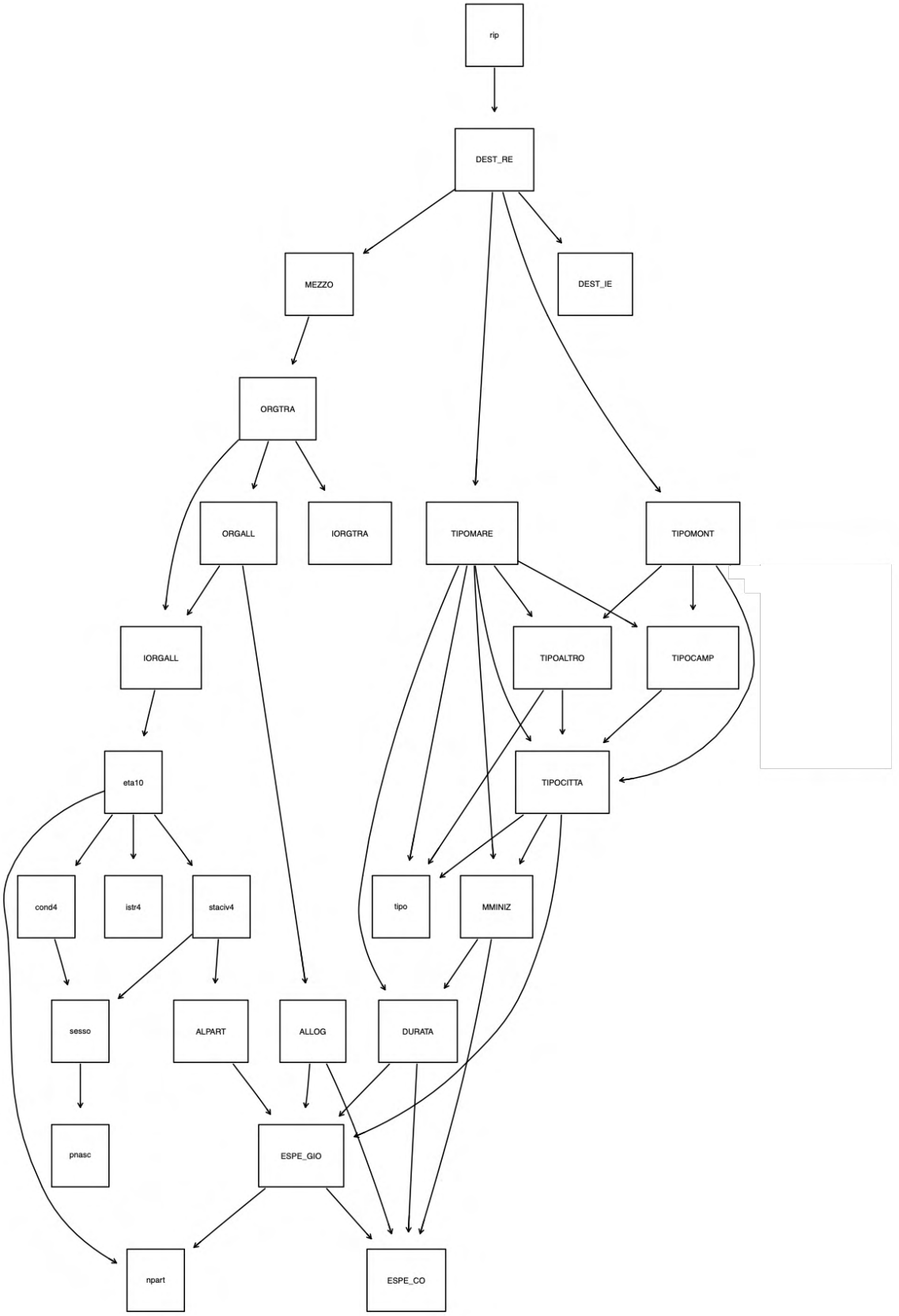
Rete Bayesiana anno 2014



Rete Bayesiana anno 2018:



Rete Bayesiana anno 2019:



Rete Bayesiana anno 2020: