



Università
Ca'Foscari
Venezia

UNIVERSITÀ CÀ FOSCARI DI VENEZIA

CORSO DI LAUREA IN AMMINISTRAZIONE FINANZA E
CONTROLLO

TESI DI LAUREA

**Le reti Bayesiane come strumento di
supporto alle decisioni aziendali: il caso
della soddisfazione clienti**

Relatore:
Ch. Prof.ssa Debora SLANZI

Laureando:
Marco Boz
Matricola: 842024

Anno Accademico:
2016 / 2017

Indice

1	Introduzione	1
1.1	Approccio alle decisioni e ai problemi aziendali	1
1.2	Soddisfazione clienti - Cenni	4
1.3	Struttura della tesi	6
2	Reti Bayesiane	8
2.1	Teoria delle probabilità - cenni	8
2.2	Introduzione alle reti Bayesiane	11
2.2.1	Componente qualitativa - La struttura: il grafo della rete	12
2.2.2	Componente quantitativa - I parametri: distribuzioni di probabilità	13
2.3	Dipendenza ed indipendenza nelle reti Bayesiane	14
2.4	Inferenza nelle reti Bayesiane	16
2.5	Inferenza esatta	18
2.5.1	Strutture a catena	18
2.5.2	Strutture a polialbero	19
2.5.3	Reti a connessioni multiple	22
2.5.4	Inferenza approssimata - Cenni	23
2.6	Apprendimento delle reti Bayesiane	24
2.6.1	Grafo noto: apprendimento dei parametri	24
2.6.2	Grafo non noto: apprendimento della struttura	26
2.6.3	Combinazione di approcci: conoscenze degli esperti e dati	29
2.7	Estensioni del modello	30
2.7.1	Diagrammi di influenza	30
2.7.2	Dynamic Bayesian Networks (DBNs)	32
2.7.3	Object Oriented Bayesian Network (OOBNs)	33
2.8	Vantaggi	35
3	Le reti Bayesiane per l'analisi della soddisfazione dei cittadini	36
3.1	Raccolta dei dati	36
3.2	Tipologia dei dati	37
3.2.1	Operazioni preliminari sui dati	37
3.2.2	Dati anagrafici	39
3.2.3	Informazione sul servizio	42
3.2.4	Ruolo operatori	43
3.2.5	Valutazione del servizio	45
3.2.6	Conoscenza ed utilizzo dei servizi complementari	48
3.2.7	Livello di contribuzione	48
3.2.8	Raccolta con modalità porta a porta nel centro storico	49

3.2.9	Analisi delle correlazioni	51
3.2.10	Alcune considerazioni finali	53
3.3	Apprendimento della struttura	53
3.3.1	Analisi degli archi individuati	54
3.3.2	Sfera d'influenza delle soddisfazioni	59
3.4	Apprendimento parametri	66
3.5	Processo d'inferenza	67
3.5.1	Identificazione dei driver di soddisfazione	68
3.5.2	Strategie aziendali e analisi di scenario	91
Conclusioni		99
Bibliografia		101
Acronimi		103
Simboli		104
Elenco delle figure		105
Elenco delle tabelle		110
Appendice		111
Appendice A: Variabili - Significato		112
Appendice B: Questionario		115
Appendice C: Software		120
Ringraziamenti		121

Capitolo 1

Introduzione

"Analysis of data is a process of inspecting, cleaning, transforming and modeling data with the goal of highlighting useful information, suggesting conclusions and supporting decision making" - Wikipedia, July 2013.

1.1 Approccio alle decisioni e ai problemi aziendali

Da sempre, ogni giorno, i *manager* delle aziende sono chiamati a compiere delle scelte per determinare le modalità ed i contenuti delle attività che la società, o le aree, che dirigono dovranno svolgere con lo scopo di perseguire efficacemente i propri obiettivi. L'elevato interesse che essi nutrono verso il **processo decisionale** è dovuto al fatto che ogni scelta intrapresa determina i risultati della gestione, risultati sulla base dei quali saranno valutati.

Ogni decisione è presa al termine di un processo che, a grandi linee, consiste nella valutazione delle possibili alternative e si conclude con la scelta della soluzione che apporti il beneficio maggiore all'impresa. Le decisioni più semplici sono prese, generalmente, facendo affidamento alla sola esperienza dei *manager*, mentre il progressivo aumentare della complessità del problema richiede necessariamente un maggior numero di elementi di cui disporre per selezionare l'alternativa migliore. In questa tesi, la fase di *decision making* è sviluppata seguendo l'approccio per fasi¹(Simon, 1960), il quale prevede che la decisione sia assunta seguendo una procedura suddivisa in cinque fasi come di seguito descritte e come rappresentato in Figura 1.1:

- (1) Analisi del problema (*Intelligence*): definizione del problema da risolvere e individuazione di tutte le informazioni ritenute utili per giungere ad una decisione;
- (2) Ricerca di possibili soluzioni (*Design*): selezione delle possibili linee d'azione capaci di fronteggiare il problema;
- (3) Valutazione e scelta dell'alternativa migliore (*Choice*): scelta della soluzione migliore tra quelle individuate nello *step* precedente sulla base dei parametri considerati più opportuni;

¹Esiste anche un secondo approccio definito "globale" che si limita a descrivere il comportamento effettivo degli individui, sostenendo che essi basano le proprie decisioni sull'esperienze passate, ragionamenti per analogia ed elementi intuitivi.

-
- (4) Attuazione della decisione (*Implementation*): attuazione operativa della decisione e conseguente attivazione delle azioni complementari necessarie;
 - (5) Controllo dei risultati ed eventuale modifica della scelta (*Control and Review*): controllo dei risultati prodotti dalla scelta con la facoltà, quando possibile, di innescare un nuovo processo decisionale nel caso i risultati non corrispondano alle aspettative o comunque non siano giudicati soddisfacenti.

L'approccio appena descritto, elenca in modo ottimale le fasi logiche del processo decisionale qualora sia svolto sotto l'ipotesi di razionalità², assunto di base che risulta essere troppo vincolante perché non tiene conto dell'aspetto psicologico e sociale del decisore. Le successive considerazioni sono riconducibili all'approccio appena descritto.

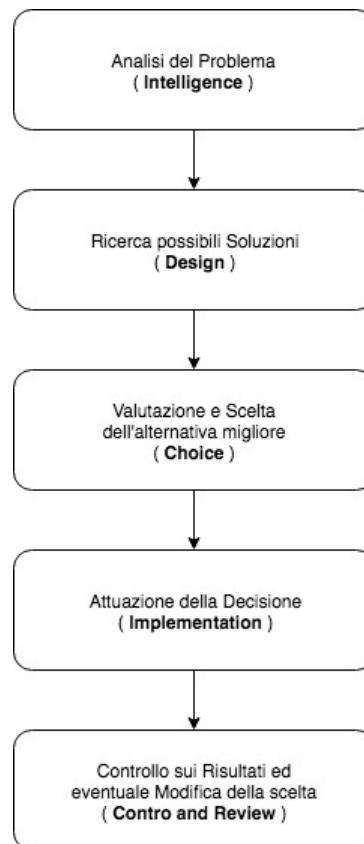


Figura 1.1: Le cinque fasi del processo decisionale

Nell'affrontare l'analisi di qualsiasi problema, l'uomo deve tener conto che le proprie abilità cognitive, mnemoniche e di ragionamento hanno dei limiti ai quali bisogna sommare l'effetto di eventuali fattori culturali, psicologici, fisici che possono introdurre ulteriori distorsioni nell'analisi (Pourret et al., 2008). I **modelli** sono nati come strumenti per gestire la realtà in maniera più semplice, nascondendone la complessità. Essi sono in grado di produrre informazioni utili per effettuare valutazioni, compiere decisioni o intraprendere azioni, a partire da un *set* di dati iniziali attraverso il processo di inferenza (Pourret et al., 2008).

Nel modellare sistemi reali non è sempre possibile disporre di dati certi, perciò bisogna

²Inizialmente era richiesta una "razionalità assoluta", successivamente è stata adottata una visione allargata del processo che richiede una "razionalità limitata". Per approfondimenti si veda (Simon, 1960).

ricorrere a modelli che siano in grado di tener conto delle incertezze associati alle informazioni raccolte. La famiglia dei modelli **probabilistici** rappresenta sistemi affetti da incertezza, la quale è quantificata attraverso l'attribuzione di una misura di probabilità ad ogni legame tra le variabili del sistema. Le reti Bayesiane appartengono proprio a questo gruppo di modelli. Korb and Nicholson (2011, pp.139) affermano che "il processo di costruzione di una rete Bayesiana possa portare ad una migliore comprensione del fenomeno oggetto d'analisi così come possa fornire uno strumento utile di supporto alle decisioni".

Le reti Bayesiane sono uno strumento molto versatile, per questo sono impiegate per studiare fenomeni di diversa natura. In passato sono già state impiegate in ambito aziendale o economico:

- (a) Soddisfazione dei clienti riguardanti dei prodotti elettronici Salini (2009);
- (b) Soddisfazione degli passeggeri dei treni in Australia Chakraborty et al. (2016);
- (c) Monitoraggio e miglioramento del servizio bancario greco Tarantola et al. (2012);
- (d) *Rating* del credito delle imprese (Pourret et al., 2008, pp. 263-277);

La questione rilevante, a questo punto, è scoprire da quali fonti reperire i dati per creare questi modelli ed ottenere informazioni utili per le decisioni strategiche.

L'analisi statistica dei dati amministrativi e contabili interni è un processo che può far emergere informazioni rilevanti, spesso non sfruttato a pieno dalle aziende. Al contrario, esse hanno sviluppato un forte interesse, indotto dal mercato e dalle norme di legge, verso la produzione di statistiche pubbliche per divulgare le informazioni sulla propria situazione economico-finanziaria. Le cause sono imputabili a molteplici fattori, tra cui: i problemi incontrati dallo statistico nell'interagire con professionisti senza nozioni statistiche, la formazione culturale dei manager che li induce ad essere diffidenti verso i risultati statistici, le difficoltà che si riscontrano nel compiere modellazioni e previsioni della realtà con sufficiente accuratezza (Brasini et al., 1999).

A quanto appena detto, si aggiunge che la raccolta dei dati su cui poi basare le proprie scelte strategiche è un'attività che solo poche aziende potevano permettersi di compiere considerando il costo elevato in termini sia di tempo sia di risorse impiegate (umane e non). Il progresso tecnologico ha portato le imprese a dotarsi di un sistema informatico che da un lato semplifica la gestione dell'azienda e che dall'altro è in grado di raccogliere le informazioni inerenti l'attività svolta. A questi sistemi si affiancano anche altri strumenti per la raccolta dei dati come le indagini *ad hoc*, i sondaggi, i *panel web*, ecc. Anche le aziende più piccole o con risorse limitate possono disporre di informazioni che fungano da supporto alle decisioni. Anche se l'operazione di raccolta delle informazioni si è semplificata, c'è il rischio che non compiendo alcun tipo di analisi, da questa grande quantità di dati non si giunga a nessuna conclusione utile per guidare il *management* nel prendere delle decisioni; in una situazione come questa, la raccolta delle informazioni rimane un processo fine a se stesso.

Per un'azienda che volesse approcciarsi all'analisi dei dati, in prima battuta entrano in gioco gli elementi messi a disposizione dalla **statistica descrittiva**, come ad esempio: indici di posizione, indici di variabilità, relazioni statistiche, distribuzioni, rappresentazioni grafiche, ecc. I risultati prodotti da questi strumenti consentono di rilevare, classificare, riassumere e rappresentare il contenuto del *database* e fungono come base di partenza per considerazioni più elaborate.

Un secondo approccio prevede di apprendere le caratteristiche del campione analizzato e di generalizzarle all’intera popolazione. Il **processo inferenziale**, ripreso in seguito e descritto ora solo a grandi linee, permette di compiere previsioni da utilizzare come supporto all’interno del processo decisionale. Lo strumento d’analisi proposto in questa tesi ha proprio questo scopo.

1.2 Soddisfazione clienti - Cenni

Premesso che lo scopo di questa tesi non è dare una spiegazione esaustiva della soddisfazione clienti o compiere un’analisi storica dei suoi metodi di misurazione e valutazione, dato che le reti Bayesiane possono analizzare qualsiasi fenomeno, si ritiene necessario spiegare, per tratti sommari, perché la *customer satisfaction* sia un elemento di fondamentale importanza per le imprese.

Il letteratura non esiste una definizione univoca di cosa sia la **soddisfazione clienti** (o *customer satisfaction*); ciò rappresenta un grosso limite nell’analizzare questa variabile poiché: (a) I ricercatori la definiscono di volta in volta in base al contesto nel quale si svolge l’analisi; (b) Risulta difficile ottenere misurazioni valide sulla soddisfazione; (c) Non è possibile confrontare o interpretare i risultati con quelli di altre ricerche per le ragioni appena elencate. Tuttavia, esaminando le definizioni nel loro complesso, emergono sempre tre elementi in comune che caratterizzano la soddisfazione dei clienti (Cote and Giese, 2002):

- (I) Si tratta della sommatoria di più risposte che dipendono dalla sfera emotiva e cognitiva, o razionale, del cliente;
- (II) Le risposte riguardano un *focus* specifico, caratterizzato dalle aspettative, dal prodotto, dagli standard di riferimento, dalle esperienze di consumo, ...;
- (III) Le risposte si collocano in un momento ben preciso (*timing*), dopo l’atto di consumo, al momento dell’acquisto, al momento della decisione d’acquisto, dopo una serie di esperienze precedenti,

Il problema legato alla definizione di questa variabile aziendale è fortemente discusso in letteratura³. Ai fini dell’analisi proposta in questa tesi, la soddisfazione clienti è definita come un giudizio valutativo globale sull’uso/consumo di un prodotto o servizio (Westbrook, 1987, p.260).

Le aziende hanno cominciato ad interessarsi alla *customer satisfaction* da quando la letteratura (Busacca and Valdani, 1999) ha introdotto la **customer based view (CBV)**, le cui proposizioni mettono in relazione tre diversi aspetti di un’impresa: (a) il valore generato per i propri clienti; (b) il valore dei clienti stessi; (c) il valore del capitale economico. Sostanzialmente, ciò che viene proposto è la focalizzazione sul ruolo centrale del cliente nel processo di sviluppo del capitale economico. La creazione di valore è uno dei requisiti fondamentali per la sopravvivenza e il successo dell’impresa, il cui valore economico è legato alle potenzialità accumulate di produrre in futuro e per lungo tempo dei flussi positivo, piuttosto che essere capace di produrne solo nell’immediato (Guatri, 1992). Si guarda ai risultati economici in un’ottica di lungo periodo. Nell’approccio CBV, questo requisito è soddisfatto focalizzando i processi aziendali sul valore offerto al cliente

³Per maggiori approfondimenti si veda (Cote and Giese, 2002)

e riconoscendo la centralità della soddisfazione clienti⁴.

Il fulcro del discorso sono quindi i clienti, quando se ne parla bisogna tener presente che i *manager* sono chiamati a perseguire un duplice obiettivo: fidelizzare i clienti esistenti ed attrarre nuovi consumatori (Kenett and Salini, 2012). In un mercato in cui le aziende perdono clienti continuamente, i dirigenti tendono a focalizzare la propria attenzione sulla ricerca di nuovi acquirenti, piuttosto che focalizzare la propria attenzione per capire "quando", "perché", il "tipo" di cliente perso e soprattutto l'ammontare di "ricavi" e i conseguenti profitti connessi alla perdita di quel cliente.

Per comprendere meglio questo problema, si consideri la tipica attività di vendita di un bene o servizio. Le aziende che puntano ad ampliare il proprio portafoglio, focalizzano il processo di vendita sull'acquisire il maggior numero di nuovi acquirenti; tutto ciò si traduce, in un primo momento, in un aumento dei volumi di vendita. Il rischio è che la spinta alle vendita avvenga senza salvaguardare l'attuale base dei clienti. Questa strategia di *business* è sostenibile fin tanto che il numero di nuovi clienti acquisiti è maggiore di quelli persi.

Se ci sono dei problemi è pressoché inutile che i *manager* cerchino soluzioni per attrarre nuovi clienti senza arginare e/o risolvere le "falle" nel proprio *business* perché anche questi nuovi clienti, prima o poi, potrebbero essere indotti a non servirsi più dei beni o servizi offerti dall'azienda. La metafora che spesso si utilizza per descrivere questa situazione è "*the hole in the bucket*" (Gonda and Khan, 2010).

Le aziende più lungimiranti focalizzano l'attenzione prima sul consolidamento del proprio portafoglio clienti e dopo sull'ampliamento dello stesso perché hanno capito che il costo necessario per acquisire un nuovo cliente è maggiore di quello per mantenerne uno già acquisito. Esse s'impongono d'inserire la massimizzazione della soddisfazione dei clienti come obiettivo prioritario nelle strategie di *marketing*. Il ragionamento alla base di quest'ottica è che strategie come queste siano capaci di agire efficacemente sui *driver* influenzando significativamente il valore aziendale percepito dal cliente in modo che cresca la sua soddisfazione e di conseguenza anche i profitti dell'azienda (Busacca and Valdani, 1999).

Tenere monitorato il livello di soddisfazione è fondamentale poiché ad ogni suo incremento, cresce il legame che il cliente ha nei confronti dell'azienda erogatrice, viceversa una diminuzione implicherà un calo nella fedeltà e di conseguenza anche nei profitti. Da ciò si deduce che una cattiva gestione della soddisfazione comporta un danno di natura economica per il valore dell'azienda (Iasevoli, 2010, p.26).

Per quantificare la soddisfazione è indispensabile reperire i dati da analizzare. Il modo migliore per sapere cosa pensa e cosa interessa ad un cliente è semplicemente chiederglie-lo. Per questo, la maggior parte delle analisi sulla soddisfazione reperisce i dati iniziali direttamente dai suoi clienti attraverso indagini CATI, *web panel*, interviste dirette, ecc. Interpellare i clienti è fondamentale per esplorare e definire le cause che li portano ad essere insoddisfatti.

A differenza di altre variabili aziendali come costi, ricavi, numero di prodotti venduti, dati anagrafici dei clienti, ecc., non esiste una misura diretta della *customer satisfaction*. Il calcolo della soddisfazione avviene indirettamente combinando insieme i dati raccolti sui così detti ***driver* di soddisfazione**. I fattori che determinano il grado complessivo di soddisfazione non sono univoci, ma vengono scelti di volta in volta sulla base delle caratteristiche, sia interne (*firm specific*) che esterne, dell'azienda (ad esempio: posizione

⁴La *customer based view* è dominata da due principi che sono appena stati riassunti, inoltre, è declinabile in sei proposizioni, per approfondimenti si veda (Busacca and Valdani, 1999)

geografica, cordialità del personale, tempi di risposte ai reclami, ecc.). Inoltre, il peso corretto da attribuire a ciascuno fattore nel calcolo della soddisfazione non è noto e deve essere estrapolato dai dati raccolti o determinato da un gruppo di "esperti". Infine, per ciascun fattore deve essere individuata un'adeguata scala di valutazione (Ferrari and Manzi, 2010).

Un aspetto rilevante nella definizione dei *driver* di soddisfazione, è che essi devono esprimere un giudizio sull'intera esperienza del cliente e non solo sul risultato finale. Misurare la soddisfazione dei clienti fornisce informazioni critiche su come l'azienda offre i propri prodotti o servizi sul mercato. In questo modo, essa può capire quali siano i suoi punti di forza e di debolezza, in modo del tutto indipendente dalla decisione di acquisto o meno del cliente. Ad esempio: l'azienda commercializza un prodotto dotato di ottime caratteristiche e di un rapporto qualità/prezzo soddisfacente, ma i clienti potrebbero decidere di non ripetere l'acquisto a causa della scarsa professionalità dell'assistenza post-vendita. Se dalla raccolta dei dati non emergesse l'inefficacia del servizio di assistenza post-vendita, l'azienda potrebbe essere indotta a pensare che il problema sia del prodotto offerto, portandola ad impiegare risorse in attività inutili in questa situazione.

Lo scopo di questa tesi è dimostrare le potenzialità delle reti Bayesiane come strumento di supporto nelle decisioni aziendali. Per comprendere il motivo per cui si analizza la *customer satisfaction* e non altre variabili, è necessaria una breve anticipazione: questi modelli, per loro definizione, sono realizzati per rappresentare sistemi caratterizzati da dati incerti. A questo punto si può intuire che studiare i "costi totali" attraverso una rete Bayesiana è poco efficace, poiché si conoscono perfettamente l'entità e l'ammontare degli elementi che compongono questa variabile. La soddisfazione clienti invece non è un fenomeno deterministico e l'elemento d'incertezza emerge ricordando che essa è calcolata indirettamente come combinazione di più fattori.

1.3 Struttura della tesi

Lo scopo della tesi è dimostrare le potenzialità delle reti Bayesiane come modello grafico probabilistico per generare informazioni rilevanti da utilizzare come supporto nelle decisioni di *business*. Volendo far riferimento alla Figura 1.1, l'analisi proposta si colloca all'interno della prima fase, in cui si analizza e si recuperano informazioni sul fenomeno, assumendo che i "dati grezzi" siano raccolti da un soggetto esterno, motivo per cui non sono presentate le modalità alternative e le relative criticità, con cui un'azienda può ottenere informazioni.

La tesi dunque, è strutturata nella seguente maniera.

Nel capitolo 2 è contenuta una panoramica di tutte le nozioni teoriche legate alle reti Bayesiane, necessarie per comprendere le operazioni che si eseguiranno e le relative conclusioni.

Il capitolo 3 è dedicato all'applicazione delle reti Bayesiane per modellare ed analizzare la soddisfazione sul servizio di raccolta rifiuti dei cittadini di un comune. In apertura sarà descritto il caso di studio, seguito dalla presentazione delle informazioni raccolte insieme alle relative operazioni necessarie per trasformarli in dati analizzabili. La rete Bayesiana utilizzata sarà descritta al termine delle sezioni 3.3 e 3.4, in cui si procede con l'apprendimento rispettivamente prima della topologia e poi dei parametri. Il cuore dell'analisi è contenuto nella sezione 3.5, dove si presenterà il processo d'inferenza sulla rete con un duplice scopo: determinare i *driver* di soddisfazione del servizio di raccolta

rifiuti e valutare le probabili conseguenze connesse alle strategie aziendali alternative tra cui i *manager* sono chiamati a scegliere.

Infine si presenteranno i risultati e le relative conclusioni.

Capitolo 2

Reti Bayesiane

I **modelli grafici** sono modelli statistici che combinano le proprietà della teoria della probabilità con quella della teoria dei grafi. In letteratura è noto come questi modelli risultino particolarmente adatti nell'affrontare contemporaneamente problemi derivanti da incertezza e complessità.

La **probabilità**, infatti, fornisce una misura dell'incertezza associata ai dati osservati mentre la teoria dei **graфи** consente di creare una struttura dati coerente con il contesto analizzato e fornisce un'interfaccia intuitiva attraverso cui i ricercatori di molte discipline possono interagire con il problema oggetto di studi (Margaritis, 2003).

Generalmente, si dispone di un modello M , descritto dall'insieme delle variabili X , e delle dipendenze dirette, V , esistenti tra coppie di elementi di X . Il modello di rete Bayesiana è rappresentato da un grafo $G \triangleq G_M(X, V)$ dove M e G sono strettamente correlati, in quanto la topologia del grafo riflette alcune delle proprietà intrinseche nel modello, come appunto le dipendenze o indipendenze tra le variabili oggetto di studio.

2.1 Teoria delle probabilità - cenni

La parte quantitativa di un modello grafico è formata dall'insieme delle probabilità che misurano l'incertezza in un sistema. Di seguito alcuni richiami alla teoria della probabilità che risulteranno utili nel contesto dei modelli descritti ed utilizzati in questa tesi. I concetti primitivi di questa disciplina sono: la prova, l'evento e la probabilità. La relazione logico-formale che li lega è contenuta nella frase (Borra and Di Ciaccio, 2008):

"In una data prova, l'evento A si verifica con la probabilità $P(A)$ ".

Per ognuno di questi elementi si possono introdurre la definizione e le proprietà che li caratterizzano.

Definizione 2.1 (Prova). *La prova (o esperimento aleatorio) è un esperimento che ha due o più possibili risultati e in cui c'è un certo grado di incertezza su quale dei risultati si presenterà.*

Ogni prova può essere suddivisa in diverse fasi che prendono il nome di *sottoprove*. Parlando di eventi bisogna tenere presente che essi possono essere appartenere a due diverse categorie:

Definizione 2.2 (Evento elementare). *Per evento elementare, indicato con ω , si intende uno dei possibili risultati di una prova.*

Definizione 2.3 (Evento non-elementare). *Per evento non-elementare A , si intende un evento che può essere a sua volta scomposto in più eventi elementari.*

In questa tesi, per questioni di semplicità, si parlerà di eventi senza compiere alcuna distinzione. Spesso è necessario introdurre anche altri insiemi di eventi, determinati come risultato di operazioni tra eventi elementari. Per analizzare questo tipo di relazioni bisogna utilizzare la teoria dell'algebra tra eventi chiamata **algebra di Boole** (Borra and Di Ciaccio, 2008, pag. 199).

Definizione 2.4 (Spazio campionario). *L'insieme di tutti i possibili eventi elementari ω viene chiamato spazio campionario ed è indicato con il simbolo Ω .*

A questo punto, sono noti tutti gli elementi necessari per la definizione del concetto di probabilità. Per riassumere quanto detto: ad una generica prova è associato uno specifico spazio campionario (Ω) formato da una collezione di n eventi costruiti secondo le leggi dell'algebra di Boole.

Definizione 2.5 (Probabilità). *La probabilità è una funzione di insieme che associa ad ogni evento $A \in \Omega$ un numero reale nell'intervallo $[0,1]$. La probabilità è indicata con $P(A)$.*

Questa definizione, anche se generica, prevede che ad ogni evento A sia associato un numero reale appartenente all'intervallo finito $[0,1]$, dove '0' indica la probabilità dell'*evento impossibile* ovvero che non accadrà mai ed '1', al contrario, la probabilità dell'*evento certo*.

Il calcolo delle probabilità deve obbedire a tre assiomi basilari, ovvero assunti che non devono essere dimostrati (Korb and Nicholson, 2011):

- Per convenzione, poiché contiene tutti i possibili eventi, allo spazio campionario è associata la massima probabilità.
Assioma 1: $P(\Omega) = 1$.
- La probabilità di un qualsiasi evento A non può mai essere negativa.
Assioma 2: $\forall A \subseteq \Omega, P(A) \geq 0$.
- Presi due eventi, A e B , tra loro esclusivi, la probabilità della loro unione è data dalla somma delle singole probabilità.
Assioma 3: $\forall A, B \subseteq \Omega, \text{if } A \cap B = \emptyset, \text{then } P(A \cup B) = P(A) + P(B)$.

Dalla definizione di questi postulati derivano numerose proprietà. Per esempio:

- La probabilità dell'insieme vuoto è: $P(\emptyset) = 0$.
- La probabilità di un evento è calcolabile a partire dalla probabilità che l'evento non si verifichi, ovvero dal suo complementare: $P(A) = 1 - P(\bar{A})$.
- Preso un evento B tale che $B \subseteq A$ la sua probabilità è: $P(B) \leq P(A)$.
- Dati due eventi qualsiasi A e B , la probabilità della loro unione (Figura 2.1c) è genericamente calcolata come: $\forall A, B \subseteq \Omega, P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

In tutte quelle situazioni in cui gli eventi sono perfettamente noti, sono equiprobabili e sono in numero finito, la probabilità può essere calcolata attraverso l'approccio classico.

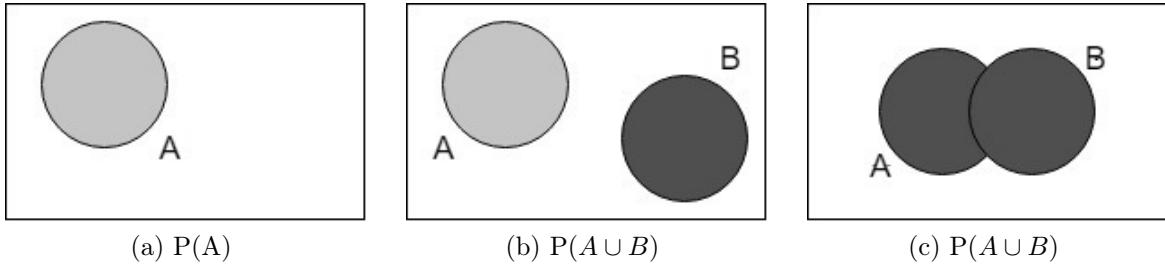


Figura 2.1: Assiomi di probabilità

Definizione 2.6 (Probabilità classica). *La probabilità di un evento A è data dal rapporto tra il numero dei casi favorevoli all'evento e il numero di casi possibili purché essi siano tutti ugualmente possibili.*

$$P(A) = \frac{n.\text{casifavorevoli}}{n.\text{casipossibili}} \quad (2.1)$$

I contesti in cui può essere usato questo metodo di calcolo sono abbastanza limitati nella pratica, in virtù del fatto che devono essere rispettate delle ipotesi di base molto restrittive. Lo strumento della probabilità condizionata costituisce le fondamenta dell'approccio Bayesiano ed è generalmente introdotta a partire dalla definizione formale (Neapolitan, 2004).

Definizione 2.7 (Probabilità condizionata). *Siano A ed B due eventi tali che $P(B) \neq 0$. La probabilità condizionata di A dato B , indicata come $P(A | B)$, è data da:*

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \quad (2.2)$$

L'utilità di questa formula consiste nel calcolare la probabilità che un primo evento A si realizzi sapendo che l'evento B si è precedentemente verificato. Dal punto di vista matematico, affinché la formula non sia indefinita è importante che la $P(B) \neq 0$. Il ragionamento sottostante parte dall'intuizione che il verificarsi di B riduce l'universo degli eventi possibili a B ; ciò pone $P(B)$ a denominatore. Il numeratore è composto considerando solamente ciò che è rimasto di A in relazione a B , cioè è rappresentato dall'intersezione dei due eventi, $P(A \cap B)$. Uno dei corollari che discendono dall'Equazione 2.2 è il principio dalle probabilità composte.

Definizione 2.8 (Principio delle probabilità composte). *Dati due eventi A e B tali che $P(A) > 0$ e $P(B) > 0$, si ha:*

$$P(A \cap B) = P(A)P(B | A) = P(B)P(A | B) \quad (2.3)$$

Il concetto esposto nell'Equazione 2.2 è il punto di partenza per stabilire quando due eventi siano tra di loro indipendenti. In particolare, due eventi A e B si dicono indipendenti quando qualsiasi condizione posta su uno di essi lascia invariata la probabilità dell'altro evento. La nozione di indipendenza è simbolicamente indicata con $A \perp B$ ed è simmetrica, quindi, $A \perp B \equiv B \perp A$ (Korb and Nicholson, 2011).

Definizione 2.9 (Indipendenza). *Due eventi A e B sono indipendenti se e solo se $A \perp B$, quindi:*

$$P(A | B) = P(A) \quad (2.4)$$

Al contrario, se due eventi sono **dipendenti** allora significa che il verificarsi di uno altera la probabilità del secondo e viceversa. L'intera filosofia della probabilità Bayesiana si fonda sull'interpretazione del teorema di Bayes, il quale permette di calcolare le probabilità condizionate.

Definizione 2.10 (Teorema di Bayes). *Dato un insieme esclusivo ed esaustivo di k eventi B_1, B_2, \dots, B_k ed un evento A , si ha:*

$$\begin{aligned} P(B_i | A) &= \frac{P(A | B_i)P(B_i)}{P(A)} \\ &= \frac{P(A | B_i)P(B_i)}{P(B_1)P(A | B_1) + P(B_2)P(A | B_2) + \dots + P(B_k)P(A | B_k)} \end{aligned} \quad (2.5)$$

Per insieme esclusivo ed esaustivo di eventi s'intende un partizionamento dello spazio campionario Ω in k eventi tali che: (i) $B_i \cap B_j = \emptyset$, $\forall i \neq j$, ossia a due a due incompatibili. (ii) $\sum_{i=1}^k B_i = \Omega$, ossia la loro unione costituisce lo spazio campionario.

Nell'Equazione 2.5 sono presenti diversi tipi di probabilità che è utile distinguere:

1. Probabilità a priori: sono le probabilità $P(B_i)$ dei singoli eventi B_i per $i = 1, 2, \dots, k$.
2. Verosimiglianze: sono le probabilità condizionate $P(A | B_i)$ per $i = 1, 2, \dots, k$.
3. Probabilità a posteriori (o *belief*): sono le probabilità condizionate $P(B_i | A)$ per $i = 1, 2, \dots, k$ che si riferiscono agli eventi B_i dopo che si è verificato A .

Questi ultimi concetti saranno ripresi successivamente parlando di inferenza probabilistica nelle reti Bayesiane.

2.2 Introduzione alle reti Bayesiane

Le reti Bayesiane sono una classe di modelli grafici probabilistici utilizzati per descrivere ed analizzare situazioni in condizioni di incertezza. La struttura di una rete Bayesiana è definita a partire da due componenti principali: i nodi, che rappresentano le variabili casuali, gli archi diretti che evidenziano le dipendenze probabilistiche tra le variabili (Korb and Nicholson, 2011). Nello specifico, una rete Bayesiana, spesso indicata con BN dall'inglese *Bayesian network*, è definita nel seguente modo (Pourret et al., 2008):

Definizione 2.11 (Rete Bayesiana). *Si considerino n variabili casuali X_1, X_2, \dots, X_n , un grafo aciclico diretto con n nodi numerati e si supponga che il nodo i del grafo sia associato alla variabile X_i . Il grafo è una rete Bayesiana, che rappresenta le variabili X_1, X_2, \dots, X_n , se:*

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | Pa(X_i)), \quad (2.6)$$

dove $Pa(X_i)$ denota l'insieme dei genitori del nodo X_i , ovvero tutte le variabili tali che nel grafo esita un arco diretto da ciascuna di queste al nodo i .

Una rete Bayesiana è generalmente indicata con $BN = (G, P)$ dove G rappresenta la struttura e P la distribuzione di probabilità.

Per comprendere a pieno il significato di questa definizione è necessario approfondire alcuni elementi.

2.2.1 Componente qualitativa - La struttura: il grafo della rete

Una rete bayesiana viene identificata a partire dalla definizione della sua struttura in termini di grafo (Krieg, 2001).

Definizione 2.12 (Grafo). *Un grafo, indicato con $G(N,A)$, è un set di nodi N , o vertici, connessi da un insieme di archi A che possono essere o meno direzionati.*

La struttura viene spesso indicata con $G = (N, A)$ ed è definita da un insieme di **nodi**, N , rappresentante le variabili casuali da analizzare, ed un insieme di **archi**, A , indicanti le relazioni di dipendenza tra i nodi del grafo. Poiché l'insieme N rappresenta esattamente l'insieme X delle variabili del sistema, nel contesto delle BN spesso la nozione assume la forma $G = (X, A)$.



Figura 2.2: Relazione diretta

Bisogna interpretare con cautela le relazioni di dipendenza, sia dirette che indirette. Intuitivamente la presenza di un arco tra due nodi indica una relazione diretta tra le variabili corrispondenti mentre l'assenza di un collegamento significa che le variabili, all'interno di questo specifico modello, sono considerate indipendenti. Si noti che è stato utilizzato il termine "relazioni di dipendenza" invece che "relazioni causali"; questa precisazione è indispensabile poiché la causalità¹ è difficile da giustificare nella maggior parte dei casi (Scutari and Denis, 2015).

Nel modello, le variabili casuali sono identificate dall'etichetta del nodo corrispondente e possono essere categoriali, discrete o continue. Le prime sono descritte all'interno di insieme contenente un numero finito di elementi mentre le variabili continue possono assumere uno degli infiniti valori compresi all'interno del proprio dominio. In ogni istante, un generico nodo della rete, X_i , può assumere in modo mutualmente esclusivo uno degli stati appartenenti al dominio in cui è definito. Nel seguito della tesi verranno considerate solo variabili categoriali o discrete, caratterizzate cioè da un numero finito di valori.

La struttura di una rete Bayesiana utilizza la metafora della famiglia per descrivere le relazioni gerarchiche tra gli elementi del grafo: un nodo è detto **genitore** (*parent*) di un **figlio** (*child*) se esiste un arco diretto che collega il primo al secondo. Un'estensione di questa terminologia identifica, prendendo un generico nodo X_i , l'insieme dei **discendenti** (*descendants*) come tutti i nodi che possono essere raggiunti attraverso un percorso diretto partendo da X_i , e gli **antenati** (*ancestor*), come l'insieme formato dai nodi da cui si può raggiungere X_i attraverso un percorso diretto (Faltin et al., 2007). Un nodo senza genitori è definito **radice** (*root*) mentre un nodo senza figli prende il nome di **foglia** (*leaf*); ogni altro nodo è classificato come **intermedio** (*intermediate*).

L'unico vincolo strutturale impone che all'interno della rete non ci siano cicli diretti,

¹Le condizioni che permettono ad una rete Bayesiana di rappresentare le causalità tra i dati sono descritte in (Pearl, 2009).

ovvero che non sia possibile partire da un nodo e ritornarvi semplicemente seguendo la direzione degli archi (Korb and Nicholson, 2011). Per questo motivo le BNs appartengano alla categoria dei **grafi aciclici diretti** o *directed acyclic graph (DAG)*². Il vincolo dell'aciclicità è necessario, poiché:

- la probabilità congiunta non sarebbe fattorizzabile come prodotto di probabilità condizionate in presenza di cicli.
- qualunque sia il numero e la natura delle dipendenze tra le variabili, esiste almeno una struttura aciclica adatta a rappresentare l'oggetto (Pourret et al., 2008).
- si garantisce che nessun nodo possa essere il suo stesso ascendente o discendente.

A differenza dei grafi non diretti, i DAG riescono a rappresentare, in modo molto flessibile, un'ampia varietà di indipendenze probabilistiche (Krieg, 2001).

2.2.2 Componente quantitativa - I parametri: distribuzioni di probabilità

Le reti Bayesiane sono annoverate tra i modelli probabilistici in virtù del fatto che le relazioni intercorrenti tra le variabili sono quantificate specificando, per ogni nodo, una distribuzione di probabilità condizionata (Korb and Nicholson, 2011). Queste distribuzioni sono rappresentate attraverso tabelle che prendono il nome di **tabella di probabilità condizionata** o *conditional probability table (CPT)* definite come segue (Krieg, 2001).

Definizione 2.13 (Tabella di probabilità condizionata). *Per ogni variabile X_i , con n nodi genitori (Y_1, Y_2, \dots, Y_n), la CPT è indicata con $P(X_i|Y_1, Y_2, \dots, Y_n)$ e contiene la probabilità associata ad ogni possibile combinazione tra gli stati di X_i e di tutti i suoi genitori.*

La definizione della tabella di probabilità avviene in modi diversi a seconda del tipo di nodo:

- Intermedio o foglia: per ogni combinazione degli stati dei genitori di un generico nodo X_i , la tabella indica la probabilità condizionata che X_i assuma uno dei valori contenuti nel proprio dominio.
- Radice: non avendo genitori, la tabella rappresenta la probabilità a priori associata ad ogni stato assunto dalla variabile. Questa è una probabilità marginale e non condizionata, poiché l'insieme dei genitori di un nodo radice è vuoto.

Utilizzando la **regola del prodotto** (*chain rule*) è possibile determinare la **distribuzione di probabilità congiunta** (*joint probability distribution*) dei nodi dell'intera BN. Essa è calcolata come prodotto delle probabilità condizionate e marginali di tutti i nodi (Krieg, 2001).

Definizione 2.14 (Distribuzione di probabilità congiunta). *Per una rete Bayesiana definita sull'insieme delle variabili $X = (X_1, X_2, \dots, X_n)$ la distribuzione di probabilità congiunta delle rete è definita come:*

$$P(X) = P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i|Pa(X_i)) \quad (2.7)$$

²A differenza delle *Markov network* nelle quali gli archi non sono direzionati.

dove $Pa(X_i)$ indica l'insieme dei genitori del nodo X_i .

La distribuzione di probabilità congiunta può essere dunque fattorizzata e scomposta nelle singole distribuzioni di probabilità locale, ognuna delle quali coinvolge un nodo della rete e l'insieme dei suoi genitori.

Dal punto di vista dell'efficienza computazionale, all'aumentare del numero dei genitori di un nodo, cresce anche la dimensione della tabella di probabilità condizionata e di conseguenza la potenza di calcolo richiesta per l'analisi. Ad esempio: in una rete booleana, dove ciascun nodo può assumere al massimo due valori, ad una variabile con n genitori è associata una TPC con 2^{n+1} probabilità.

2.3 Dipendenza ed indipendenza nelle reti Bayesiane

Caratteristica fondamentale di una BN è la capacità di catturare e rappresentare le relazioni che intercorrono tra le variabili del sistema analizzato. Per questo motivo è indispensabile far chiarezza sui concetti rispettivamente di dipendenza e indipendenza al loro interno. Il concetto di *d-separation* viene introdotto come criterio grafico per identificare le indipendenze che sussistono tra le variabili data la struttura della rete. Per comprenderne al meglio il significato è necessario definire la nozione di percorso bloccato. Per far ciò bisogna definire le possibili relazioni all'interno di reti definite su insiemi di tre variabili (Krieg, 2001):

1. **Connessioni seriali** (*serial connections*): nella Figura 2.3a è rappresentata una rete formata da tre nodi in cui X influisce su Y che a sua volta influisce su Z . In questa situazione si crea un'indipendenza condizionata, poiché:

$$P(Z | X \wedge Y) = P(Z | Y) \equiv X \perp Z | Y$$

Quest'uguaglianza indica che la probabilità di Z sapendo che X e Y si sono verificati è uguale a quella che si otterrebbe sapendo che solo Y si è verificato; inizializzando Y ad un qualche valore, conoscere X non ha alcuna influenza sulla probabilità di Z . Questa relazione di indipendenza condizionata è indicata come $X \perp Z | Y$.

2. **Connessioni divergenti** (*diverging connections*): in una *v-structure* (Figura 2.3b), il processo di inferenza può essere compiuto o su un qualsiasi nodo figlio conoscendo lo stato del genitore o sul *parent* conoscendo lo stato di uno dei nodi *child*. Nel caso in cui sia data un'evidenza sul genitore, il flusso delle informazioni tra i nodi figli è bloccato e la probabilità dei singoli stati di un nodo figlio non è influenzata in alcun modo dalle evidenze riscontrate per gli altri figli:

$$P(Z|X \wedge Y) = P(Z|Y) \equiv X \perp Z | Y$$

Conoscere lo stato del genitore Y instaura dunque una relazione di indipendenza condizionata tra i figli X e Z .

3. **Connessioni Convergenti** (*converging connections*): la struttura ad effetti comuni, rappresentata nella Figura 2.3c, descrive la situazione in cui più variabili producano uno stesso effetto. I genitori sono marginalmente dipendenti se non si hanno evidenze sul figlio mentre in caso contrario diventano condizionalmente indipendenti (o *d-connected*):

$$P(Z|X \wedge Y) \neq P(Z|Y) \equiv X \not\perp Z | Y$$

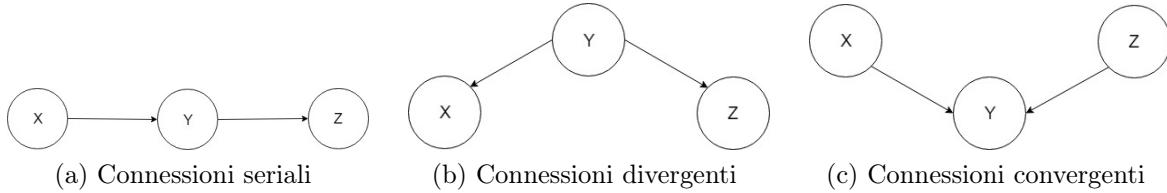


Figura 2.3: Casi di indipendenza condizionale

Dopo aver analizzato questi tre tipi di connessione, si può definire il concetto di d-separazione, elemento indispensabile per la definizione e l'identificazione delle indipendenze in una rete Bayesiana (Pearl, 1995).

Definizione 2.15 (D-separazione). *I nodi X e Z si dicono d-separati se in ogni sequenza diretta di archi (cammino) tra X e Z esiste un nodo Y tale che:*

1. *Y sia noto e sia in connessione seriale o divergente;*
2. *Y è in connessione convergente e nè Y nè nessuno dei suoi discendenti è noto.*

Al contrario, due nodi X ed Z sono **d-connessi** dato Y ($X \not\perp\!\!\!\perp Z|Y$) se esiste un percorso dal nodo X al nodo Z che non sia bloccato dato Y . Alla base dell'indipendenza tra i nodi ci sono tre assunzioni indispensabili (Margaritis, 2003):

Definizione 2.16 (Sufficienza causale). *Non esistono, nel dominio, variabili non osservate che siano genitori di uno o più variabili osservate.*

Il senso di questa prima assunzione è vietare di poter spiegare, utilizzando variabili che non siano state osservate, eventuali relazioni di indipendenza emergenti dai dati o, all'opposto, una loro mancanza.

Definizione 2.17 (Proprietà di Markov). *Data una rete bayesiana BN, ogni variabile è indipendente da tutti i suoi non discendenti dati i suoi genitori.*

Questa proprietà prevede che non ci siano dipendenze dirette nel sistema che non siano rappresentate da un arco.

Definizione 2.18 (Fedeltà). *Il grafo G di una BN e una distribuzione di probabilità P si dicono fedeli l'un con l'altro se tutte e solo le relazioni di indipendenza valide in P sono le stesse rilevate dalla proprietà di Markov.*

Le reti che rispettano tale proprietà sono chiamate **independency-maps (I-map)**: in esse due nodi non collegati da un arco corrispondono a variabili indipendenti. In modo analogo a quanto appena detto, se nel grafo tutti i nodi connessi si riferiscono a variabili dipendenti nella realtà, allora si parla di **dependency-map (D-map)**. Infine, una BN che contemporaneamente abbia le caratteristiche sia di una D-map sia di una I-map e che soddisfi una serie di condizioni necessarie e sufficienti, allora è definita **perfect map** (Pearl, 1988).

La decomposizione della distribuzione di probabilità congiunta nell'Equazione 2.7 identifica nell'insieme dei genitori di ogni nodo, l'insieme delle variabili che condizionano e quindi hanno un'influenza diretta sul nodo stesso (per la Definizione 2.17). Se lo scopo dell'analisi è quello di fare inferenza sul sistema oggetto di studio, è necessario allargare l'insieme dei nodi che direttamente o indirettamente agiscono su un nodo *target* per

consentire il passaggio dell'informazione tenendo in considerazione tutte le varie tipologie di connessioni che si possono riscontrare in gruppi di nodi tra loro comunicanti. Il sottoinsieme minimo di nodi che rispecchia queste condizioni è il *Markov blanket* (*MB*).

Definizione 2.19 (Markov blanket). *Il Markov blanket di un nodo $X \in N$ è il sottoinsieme minimo S di N , tale che:*

$$X \perp_G N - S - X|S \quad (2.8)$$

Integrando nella formula appena descritta il concetto di fedeltà, si ottiene definitivamente l'insieme dei nodi ricercato.

Corollario 2.1 (Markov blanket e Fedeltà). *Assumendo l'ipotesi di fedeltà, la definizione 2.19 implica che S dia il sottoinsieme minimo di N , tale che:*

$$X \perp_P N - S - X|S \quad (2.9)$$

Il corollario 2.1 è la definizione formale di ciò che si sta cercando cioè il sottoinsieme di nodi che rende tutto il resto ridondante quando si svolge inferenza su un dato nodo (Scutari and Denis, 2015, pp. 91).

Definizione 2.20 (Composizione del Markov blanket). *Il Markov blanket di un nodo X è l'insieme che contiene i genitori di X , i figli di X e tutti gli altri nodi che condividono un figlio con X .*

Il concetto di *Markov blanket* risulterà fondamentale nel processo inferenziale poiché, introducendo l'evidenza su un nodo X , sarà sufficiente limitarsi ad elaborare le variazioni dei parametri dei nodi compresi nel *Markov balnket* di X e senza dover processare l'intera rete.

2.4 Inferenza nelle reti Bayesiane

Le reti Bayesiane sono dei modelli statistici concepiti per essere dinamici nel tempo permettendo l'integrazione di nuove informazioni relative al fenomeno oggetto di studio. Dopo aver definito la struttura della rete sull'insieme dei nodi $X = (X_1, X_2, \dots, X_n)$ e aver quantificato l'incertezza attraverso la stima dei parametri $P(X_i|Pa(X_i))$, è possibile aggiornare la componente quantitativa della rete attraverso il calcolo delle probabilità a posteriori sui nodi, cioè le probabilità aggiornate delle variabili quando si ottengono nuove informazioni sul sistema. Questo processo si definisce come **inferenza probabilistica** (o *belief updating*).

Definizione 2.21 (Evidenza). *Sia E un sottoinsieme di variabili di X che assume complessivamente lo stato e ($x_i = e_i$ con $x_i \in X$ e $e_i \in e$), e viene detta evidenza e si riferisce a qualsiasi nuova informazione sia osservata ed introdotta nel modello.*

Durante l'analisi di un fenomeno, le informazioni raccolte possono avere caratteristiche diverse che necessitano di essere tenute in considerazione e trattate in modo adeguato. Affermare che un nodo X_i sia stato osservato trovarsi in un preciso stato e_i rappresenta un'informazione certa, $P(X_i = e_i) = 1$, che prende il nome di evidenza **specifica** (*hard evidence*).

Al contrario, quando si esclude che una variabile assuma uno o più valori, si parla di

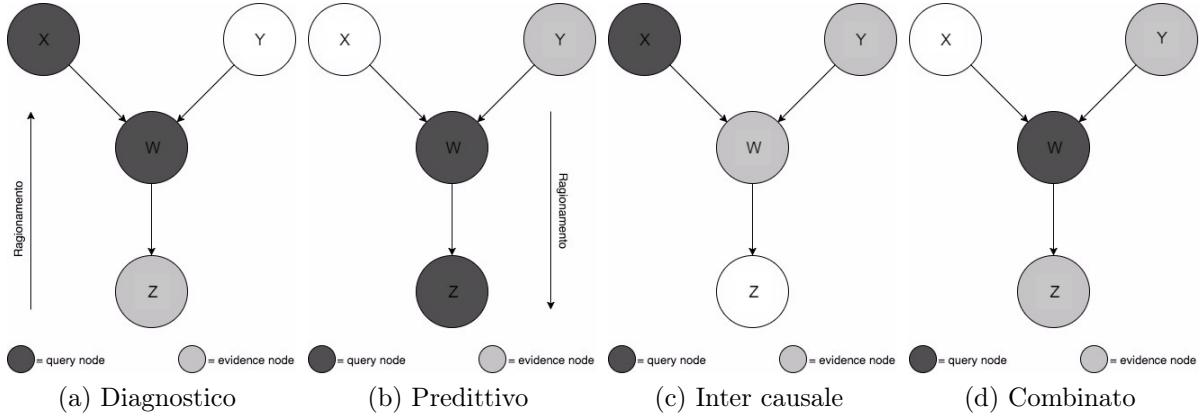


Figura 2.4: Tipologie di ragionamento

evidenza **negativa**; dire che X_i non si trova nello stato e_i equivale a dire che la $P(X_i = e_i) = 0$. Nella pratica, spesso le informazioni non sono precise e quindi incorporano un certo grado di incertezza. Quando un'osservazione non è in grado di identificare direttamente lo stato di una variabile ma può solo darne un'indicazione probabilistica, prende il nome di evidenza **virtuale** (Korb and Nicholson, 2011).

La capacità poi di poter compiere ragionamenti di diversa natura è uno dei motivi per cui le reti Bayesiane sono considerate degli strumenti flessibili. Quattro sono le logiche (Figura 2.4) utilizzabili nelle BN (Korb and Nicholson, 2011). A seconda di quali siano i nodi in cui si inserisce l'evidenza nel modello (*evidence node*) e quali invece quelli su cui compiere inferenza (*query node*), si distinguono i ragionamenti in:

- (i) **Diagnostico** (*bottom-up*): l'evidenza riguarda un effetto o sintomo (Z) e si cerca di ragionare in termini di cause (X, Y, W).
- (ii) **Predittivo** (*top-down*): pervengono nuove informazioni riguardanti le cause (X o Y) e si vuole determinare, seguendo la direzione degli archi, i nuovi parametri riguardanti gli effetti (W, Z).
- (iii) **Inter causale** (*explaining away*): logica alla base delle *v-structure* in cui si ragiona sulle cause reciproche che hanno un effetto comune; in questo caso, anche se inizialmente le cause sono indipendenti, entrare in possesso di informazioni riguardanti l'effetto (W) ed una causa (Y) permette di modificare la probabilità a posteriori associata alle altre cause (X).
- (iv) **Combinato**: ogni combinazione delle precedenti forme di ragionamento.

L'inferenza probabilistica consiste dunque nella propagazione di un'informazione su alcuni nodi all'interno della rete Bayesiana per stimarne gli effetti sulle altre variabili (Sucar, 2015).

Definizione 2.22 (Probabilità a posteriori - Belief). *Data una BN con n nodi, per ogni variabile X_i data l'evidenza e , la probabilità a posteriori che X_i assuma lo stato x_i è definita come:*

$$Bel(X_i = x_i) \triangleq P(x_i|e) \quad (2.10)$$

Il vettore contenente le probabilità a posteriori per ogni possibile stato di X_i si indica con $Bel(X_i)$ (Krieg, 2001).

La capacità di adattarsi ogni qual volta sopraggiungano nuove informazioni è sicuramente una delle caratteristiche più importanti delle reti Bayesiane. Conoscere le leggi che determinano la propagazione delle evidenze nel modello consente di calcolare le variazioni che subiranno le probabilità delle singole variabili. A seconda del tipo di struttura, esistono in letteratura metodologie per l'inferenza probabilistica esatta o approssimata. Per aggiornare le reti più semplici è sufficiente applicare ripetutamente il teorema di Bayes mentre per reti con un grado di complessità strutturale elevato esistono algoritmi approssimati per il calcolo delle probabilità a posteriori.

2.5 Inferenza esatta

La velocità e l'efficiente del processo d'inferenza dipendono da fattori come la struttura della rete, il numero di connessioni per ogni nodo, la collocazione delle evidenze e dei nodi *query*, ecc. Se il carico computazionale richiesto per l'elaborazione del passaggio di informazioni può essere gestito senza commettere errori allora l'inferenza si dice **esatta** o certa.

In letteratura esistono diversi approcci che dipendono principalmente dalla tipologia di struttura della rete. Nel seguito verranno introdotti alcuni approcci d'inferenza esatta per le principali tipologie di strutture.

2.5.1 Strutture a catena

Le catene rappresentano le strutture più semplici sulle quali compiere inferenza. L'elaborazione di questo tipo di reti consiste nell'applicazione reiterativa del teorema di Bayes.

Catena a 2 nodi Il caso più semplice è quello in cui la rete è costituita semplicemente da due nodi: $X \rightarrow Y$. Come già spiegato in precedenza, la logica con la quale si svolge l'inferenza dipende dalla posizione in cui si inserisce l'evidenza all'interno del modello.

Se l'evidenza è introdotta attraverso un nodo genitore, $X = e$, la probabilità a posteriori di Y ($Bel(Y)$) corrisponde alla probabilità condizionata $P(Y | X = e)$.

Se, invece, le nuove informazioni riguardano un nodo figlio, $Y = e$, allora il processo di aggiornamento richiede l'applicazione del teorema di Bayes (Korb and Nicholson, 2011):

$$\begin{aligned} Bel(X = x) &= P(X = x | Y = e) \\ &= \frac{P(Y = e | X = x)P(X = x)}{P(Y = e)} \\ &= \alpha P(x)\lambda(x) \end{aligned} \tag{2.11}$$

dove con $P(x)$ si indica la probabilità a priori marginale per il nodo X , $P(X = x)$, e con $\lambda = P(Y = e | X = x)$ la probabilità condizionata. Poiché la somma delle probabilità dei valori di X deve essere 1, il termine α è la costante di normalizzazione ed è calcolato come:

$$\alpha = \frac{1}{P(Y = e)} \tag{2.12}$$

Catena a 3 nodi L’aggiornamento dei parametri in una rete composta da tre nodi concatenati, $X \rightarrow Y \rightarrow Z$, è svolto in modo analogo a quello descritto precedentemente. Inserire nel modello un’evidenza riguardante il nodo $X = e$, rende necessario applicare la regola del prodotto seguendo la direzione imposta dagli archi per compiere inferenza (Korb and Nicholson, 2011).

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

Se sopraggiungono nuovi elementi riguardanti il nodo, $Z = e'$, il ragionamento diagnostico per ottenere la probabilità a posteriori di X è svolto semplicemente applicando il teorema di Bayes (Korb and Nicholson, 2011):

$$\begin{aligned} Bel(X = x) &= P(X = x | Z = e') \\ &= \frac{P(Z = e' | X = x)P(X = x)}{P(Z = e')} \\ &= \frac{\sum_{Y=y} P(Z = e' | Y = y, X = x)P(Y = y | X = x)P(X = x)}{P(Z = e')} \\ &= \frac{\sum_{Y=y} P(Z = e' | Y = y)P(Y = y | X = x)P(X = x)}{P(Z = e')} (Z \perp X | Y) \\ &= \alpha P(x)\lambda(x) \end{aligned} \tag{2.13}$$

dove con $\lambda(x)$ si indica la probabilità condizionata, calcolata come:

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

Alla luce dei risultati ottenuti, l’inferenza su reti composte da semplici strutture a catena è risolvibile in maniera chiara utilizzando la definizione di indipendenza condizionate e applicando il teorema di Bayes. Questi stessi ragionamenti costituiscono parte delle fondamenta degli algoritmi utilizzati per le reti più complesse.

2.5.2 Strutture a polialbero

Durante la progettazione di un algoritmo, uno degli obiettivi più importanti consiste nell’implementarlo in modo che consumi meno risorse possibili; in altri termini, che abbia un basso costo computazionale. Gli algoritmi creati per svolgere inferenza nelle reti bayesiane si differenziano per la diversa potenza di calcolo richiesta da ciascuno. Ad esempio, la tecnica conosciuta come "eliminazione del nodo" (*elimination node*) è computazionalmente troppo costosa, nonostante sia in grado di portare a termine con successo l’aggiornamento delle probabilità in una rete (Korb and Nicholson, 2011).

L’algoritmo pensato da Kim e Pearl, detto "*message passing*" supera questi limiti di elaborazione. Il ragionamento di base prevede di aggiornare la probabilità a posteriori di ogni nodo sulla base esclusivamente dei messaggi trasmessi dai nodi immediatamente vicini. Un’evidenza introdotta nella rete si propaga secondo le direzioni degli archi ed influenza le probabilità dei nodi vicini, a loro volta queste modifiche si propagheranno nel modello. In questo modo i calcoli per determinare la $Bel(X)$ di una generica variabile X richiedono in *input* solo i messaggi inviati dai nodi vicini detti *neighbours*. Questo

tipo di approccio viene implementato in reti con struttura a polialbero in cui tra ogni paio di nodi esiste al massimo un percorso indipendentemente dalla direzione degli archi. Per questo motivo sono anche chiamate **reti a connessioni singole** (*singly-connected networks*). La Figura 2.5 mostra un esempio di struttura a polialbero.

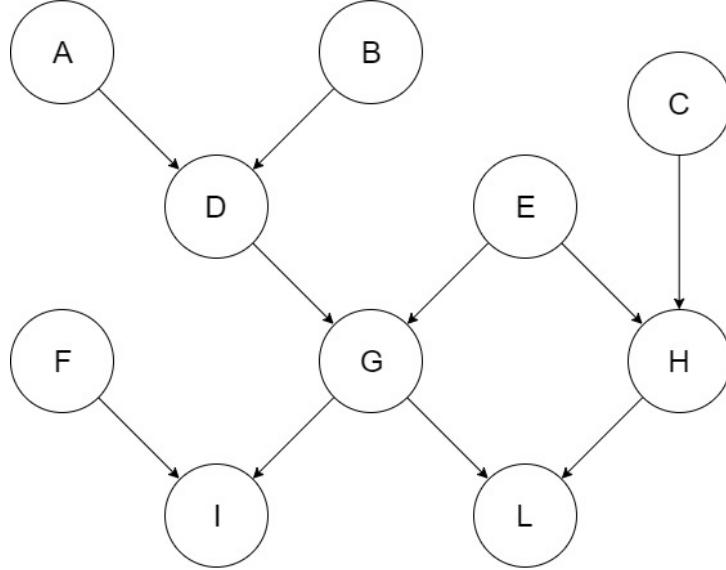


Figura 2.5: Struttura polialbero

Prendendo in esame un generico *query node* X e un set di evidenze E che non includa X , l'obiettivo è determinare la $Bel(X)$ attraverso il calcolo di $P(X | E)$. L'insieme delle informazioni contenute in E possono essere divise in due gruppi (Krieg, 2001):

- Supporto predittivo: le informazioni, contenute nel vettore $\pi(X)$, che X riceve dai nodi genitori sono indicate con e_X^+ . Se X fosse un nodo radice allora e_X^+ sarebbe composto dalle probabilità marginali a priori ad esso associate.
- Supporto diagnostico: le evidenze che si propagano dai nodi figli ad X sono indicate con e_X^- e sono racchiuse nel vettore $\lambda(X)$.

Assumendo che questi sottoinsiemi siano indipendenti, applicando il teorema di Bayes si ottiene:

$$\begin{aligned} Bel(X = x) &= P(X = x | e_X^+, e_X^-) \\ &= \frac{P(e_X^- | X = x, e_X^+) P(X = x | e_X^+)}{P(e_X^- | e_X^+)} \\ &= \frac{P(e_X^- | X = x) P(X = x | e_X^+)}{P(e_X^-)} \end{aligned}$$

Definendo $\pi(x) = P(x | e_X^+)$ e $\lambda(x) = P(e_X^- | x)$, la *belief* diventa:

$$Bel(X = x) = \alpha \pi(x) \lambda(x) \quad (2.14)$$

dove con α , anche in questo caso, si indica la costante di normalizzazione.

Nonostante il risultato sia simile a quello ottenuto nell'Equazione 2.11 trattando le strutture a catena, questa premessa è necessaria per introdurre la logica sottostante l'algoritmo di Kim e Pearl.

Kim and Pearl's message passing algorithm Le evidenze ricevute dai nodi genitori sono indicate con U_1, \dots, U_n mentre quelle in arrivo dai figli con Y_1, \dots, Y_m . Prendendo in analisi un generico nodo X del *polytree* i parametri richiesti dall'algoritmo sono tre:

- (I) il supporto predittivo a cui contribuisce ogni nodo entrante $U_i \rightarrow X$:

$$\pi_x(U_i) = P(U_i | E_{U_i/X})$$

dove $E_{U_i/X}$ incorpora l'intera evidenza trasmessa tramite U_i ad X .

- (II) il supporto diagnostico composto da ogni arco $X \rightarrow Y_j$ uscente da X :

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

dove $E_{Y_j/X}$ rappresenta l'informazione, passante per X ed inviata ai figli Y_j .

- (III) i valori della TPC, $P(X | U_1, \dots, U_n)$, che lega la variabile X ai suoi nodi genitori.

L'aggiornamento della parte quantitativa del modello è compiuto attraverso l'elaborazione di queste probabilità in tre fasi successive che possono essere svolte in qualsiasi ordine:

1. **Belief updating.** In questa fase si modifica la probabilità a posteriori di X sulla base delle informazioni ricevute dai nodi vicini, comprese le variazioni delle loro *belief*, contenute in $\lambda_{Y_j}(X)$ e $\pi_X(U_i)$. Il calcolo da svolgere è lo stesso riportato nell'Equazione 2.14:

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

i cui elementi sono definiti come segue:

$$\lambda(x_i) = \begin{cases} 1 & \text{se l'evidenza è } X = x_i \\ 0 & \text{se l'evidenza è per un altro } x_j \\ \prod_j \lambda_{Y_j}(x_i) & \text{in tutti gli altri casi} \end{cases} \quad (2.15)$$

L'Equazione 2.15 mostra come determinare gli elementi del vettore $\lambda(x_i)$: se il valore introdotto dall'evidenza è x_i allora il parametro assume valore '1', al contrario sarà pari a '0' nel caso in cui riguardi un altro valore x_j . In tutti gli altri casi, quando le nuove informazioni non riguardano la variabile X , il valore del parametro sarà dato dal prodotto di tutti i messaggi ricevuti dai figli.

$$\pi(x_i) = \sum_{u_1, \dots, u_n} P(x_i | u_1, \dots, u_n) \prod_i \pi_X(u_i) \quad (2.16)$$

Per quanto concerne i messaggi ricevuti dai nodi genitori, l'Equazione 2.16 determina i parametri di $\pi(x_i)$ come il prodotto degli elementi della probabilità condizionata e dei messaggi π ricevuti dai nodi genitori.

L'elemento α è la costante di normalizzazione che rende $\sum_{x_i} Bel(X = x_i) = 1$.

2. **Bottom-Up propagation:** In questa fase i genitori ricevono nuove informazioni dal nodo X .

$$\lambda(u_i) = \sum_{x_i} \lambda(x_i) \sum_{u_k: k \neq i} P(x_i | u_1, \dots, u_n) \prod_{k \neq i} \pi_X(u_k) \quad (2.17)$$

I messaggi λ generati da X ed inviati ad ogni genitore sono frutto della combinazione di tre tipi di informazione: quelle ricevute dai figli, quelle derivanti da tutti gli altri genitori e dai valori della probabilità condizionata.

3. **Top-Down propagation:** La terza fase è quella in cui X elabora nuovi messaggi da inviare ai suoi nodi figli.

$$\pi_{Y_j}(x_i) = \begin{cases} 1 & \text{se il valore dell'evidenza introdotta è } x_i \\ 0 & \text{se l'evidenza riguarda un altro valore } x_j \\ \alpha [\prod_{k \neq j} \lambda_{Y_k}(x_i)] \sum_{u_1, \dots, u_n} P(x_i | u_1, \dots, u_n) \prod_i \pi_X(u_i) \\ = \frac{\alpha Bel(x_i)}{\lambda_{Y_j}(x_i)} \end{cases} \quad (2.18)$$

Quanto appena descritto in termini matematici, rappresenta il criterio per stabilire il contenuto dei messaggi $\pi_{Y_j}(x_i)$ inviati da X ai suoi figli. Assumerà valore '1' se l'evidenza per il nodo X è x_i mentre sarà uguale a '0' se riguarda un altro stato x_j . Se le informazioni introdotte nella rete non interessano la variabile X , allora il messaggio trasmesso ad un generico nodo figlio Y_j sarà influenzato sia dalle informazioni provenienti dagli altri figli, sia da quelle ricevute dai genitori, sia dai valori contenuti nella probabilità condizionata.

Il corretto funzionamento di quest'algoritmo prevede che, prima dell'introduzione di qualsiasi evidenza, si compiano le seguenti operazioni:

- Inizializzare tutti i valori di λ ed i messaggi di λ e π a '1' .
- Per ogni nodo radice, se il nodo W non ha genitori allora bisogna inizializzare $\pi(W)$ alla sua probabilità a priori, $P(W)$.

2.5.3 Reti a connessioni multiple

La maggior parte dei fenomeni analizzati non possono essere modellati usando semplici strutture a catena o polialbero, bensì richiedono la costruzione di reti più complesse. Le **reti a connessioni multiple** (*multiply connected network (MCN)*) sono chiamate così perché almeno una coppia di nodi è connessa da più di un percorso. Ciò implica che una variabile può influenzarne un'altra attraverso più di un passaggio di informazioni.

Una rete come quella nella Figura 2.6 in cui i nodi X , Y , W , Z costituiscono parte

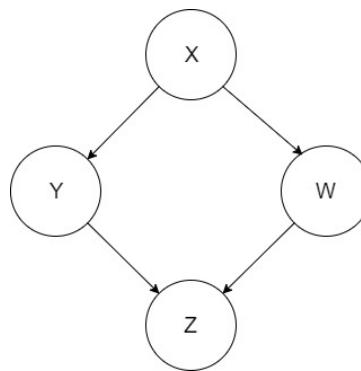


Figura 2.6: Multiply-connected network

di un ciclo non direzionale, l'algoritmo inferenziale di Kim e Pearl non potrebbe essere utilizzato. La ragione di questo limite è che l'esistenza di più percorsi tra due nodi, per esempio tra X e Z in Figura 2.6, comporta che uno stesso messaggio creato dall'evidenza sul nodo X , raggiunga il nodo Z percorrendo due strade diverse.

I metodi inferenziali basati sul *clustering*, come l'algoritmo basato sui *junction trees*, sono stati creati appositamente per elaborare reti di questa complessità.

Junction tree Clustering Algorithm Questo approccio di inferenza probabilistica risulta essere uno tra i più utilizzati ed efficienti. Appartiene alla famiglia dei metodi di *clustering*, i quali lavorano in due fasi successive: nella prima, l’obiettivo è trasformare la rete in una struttura tipo *polytree* fondendo nodi o rimuovendo i percorsi multipli attraverso cui potrebbe propagarsi l’evidenza; nella seconda si svolge il vero e proprio processo d’inferenza sulla nuova struttura. Tutto ciò è possibile perché una *multiply-connected network* è sempre trasformabile, in maniera più o meno semplice, in un *polytree*. L’algoritmo è articolato in sei fasi (Korb and Nicholson, 2011):

1. Moralizzare: connettere assieme tutti i genitori e rendere tutti gli archi indiretti; il grafo risultato è chiamato *moral graph*. Sostanzialmente, se non è già presente un arco diretto tra due nodi che hanno un figlio in comune, allora viene aggiunto un arco indiretto tra i due genitori. Per esempio, nella Figura 2.6 poiché Y e W hanno un figlio in comune Z , l’algoritmo prevede la creazione di un arco indiretto tra loro.
2. Triangolare: aggiungere archi indiretti in modo tale che ogni ciclo di lunghezza > 3 abbia una corda³. Il grafico prodotto prende il nome di *grafo triangolato*. Nella pratica, sorge il problema di trovare il metodo di triangolazione ottima, poiché l’utilizzo di metodologie differenti porta alla creazione di *cluster* differenti.
3. Creare una nuova struttura: questa fase porta alla creazione del *junction tree* e consiste nell’identificare le cricche⁴ massime⁵ e renderle i nuovi nodi compositi.
4. Creare i separatori: ad ogni arco è associato un separatore costituito dall’intersezione dei nodi adiacenti, utilizzato per inviare e ricevere i messaggi nell’ultimo *step*.
5. Calcolare i nuovi parametri: sulla base delle configurazioni delle variabili che lo compongono, ad ogni nuovo nodo X del *junction tree* è associata una nuova tabella di probabilità condizionata, costruita con i seguenti criteri:
 - (a) si prenda una *clique* Y nel grafo che contenga X e tutti i suoi nodi genitori.
 - (b) si moltipichi $P(X | Pa(X))$ nella tabella di Y .
6. *Belief updating*: inserimento nel modello dell’evidenza e aggiornamento dei parametri attraverso un algoritmo di propagazione dei messaggi come quello di Kim e Pearl.

Qualsiasi struttura ottenuta al termine di questo procedimento è in grado di rappresentare la stessa distribuzione di probabilità congiunta.

2.5.4 Inferenza approssimata - Cenni

Diametralmente opposta all’inferenza certa descritta nel paragrafo precedente, l’inferenza **approssimata** entra in gioco quando il carico computazionale richiesto è troppo elevato.

³Corda: in un ciclo di lunghezza n si definisce corda un arco tra una coppia di nodi non adiacenti.

⁴Cricca (o *clique*): nella teoria dei grafi è un insieme V di vertici in un grafo non orientato G , tale che, per ogni coppia di vertici in V , esiste un arco che li collega.

⁵Cricca massima: nella teoria dei grafi è un sottografo che non può essere ingrandito aggiungendo un altro nodo.

Questo si verifica nel caso di BN complesse o densamente connesse. In questa situazione è necessario ricorrere ad algoritmi che calcolino i nuovi parametri mediante delle approssimazioni.

Gli algoritmi ideati per compiere questo tipo di elaborazioni sono molteplici ed ognuno ha caratteristiche diverse. Possono basarsi sullo svolgimento di simulazioni (*stochastic simulation*), far affidamento sulla rappresentazione grafica, sull'utilizzo di evidenze virtuali, su valutazioni approssimative dell'inferenza. A titolo esemplificativo se ne elencano alcuni: *logic sampling*, *likelihood weighting*, *Markov chain Monte Carlo*.

Per maggiori approfondimenti si veda: (Korb and Nicholson, 2011, pag.94), (Sucar, 2015).

2.6 Apprendimento delle reti Bayesiane

In molti casi il modello di rete non è noto ed è necessario stimarlo attraverso i dati di cui si è in possesso. I dati in questione possono provenire da fonti diverse: banche dati, esperti del fenomeno analizzato (*expert knowledge*), esperimenti controllati, questionari (*survey*), ecc. Per apprendere una rete sono necessari due elementi: i parametri associati ad ogni nodo e la topologia della BN.

2.6.1 Grafo noto: apprendimento dei parametri

Supponendo che la struttura sia nota, i parametri sono appresi stimando le tabelle di probabilità delle variabili sulla base dei dati raccolti e delle relazioni di dipendenza e indipendenza individuate nel grafo. Gli ostacoli principali nella stima sono l'eventuale assenza, incertezza od inosservabilità dei dati relativi ad una o più variabili.

Caso 1: Dati completi In questa situazione la topologia della rete è nota e i dati sono completi e sufficienti. I parametri sono stimati con il metodo della *massima verosomiglianza* o *maximum likelihood* (ML) che costruisce la tabella di probabilità condizionata sulla base delle frequenze di ogni valore o combinazione di valori.

Per esempio: considerando una struttura convergente come quella in Figura 2.3c, in cui il nodo Y ha due genitori X e Z , la distribuzione di Y è calcolata nel seguente modo:

$$P(Y = y_i \mid X = x_j, Z = z_k) = N_{ijk}/N_{jk}$$

dove N_{ijk} è il numero di volte che $Y = Y_i$, $X = X_j$ e $Z = Z_k$ e N_{jk} è il numero totale di casi in cui si verifica che $X = X_j$ e $Z = Z_k$.

Caso 2: Dati incompleti Analizzare problemi complessi spesso richiede di elaborare informazioni di scarsa qualità o in qualche modo corrotte. Il termine "incompleto", in questo contesto, è utilizzato in una concezione più ampia ed incorpora tutti quei casi in cui i dati siano incompleti, incerti o mancanti.

Un primo ostacolo si verifica quando nel database un particolare evento A non si verifica mai. Un evento di questo tipo viene generalmente classificato come impossibile e la sua frequenza è nulla con probabilità, di conseguenza, pari a 0. Il punto della questione è che non è detto che A sia realmente impossibile e non si possa verificare in futuro. Questa situazione può essere evitata utilizzando un metodo di *smoothing* il cui obiettivo è di eliminare le probabilità uguali a zero. Una delle tecniche più semplici ed utilizzate è il

Laplacian smoothing (Sucar, 2015) che consiste nell'inizializzare le probabilità con una distribuzione uniforme per poi aggiornarle attraverso i dati. Considerando una variabile discreta X con k possibili valori, inizialmente ogni probabilità è settata a $P(x_i) = 1/k$, ovvero ciascun valore risulta essere equiprobabile. Prendendo poi N osservazioni che identificano il database, in cui ogni valore x_i ricorre m volte, si stima la probabilità come segue:

$$P(x_i) = \frac{1 + m}{k + N}$$

Spesso nella pratica ai parametri è associato un certo grado di incertezza. Il problema, generalmente, è risolto modellando quest'incertezza con una distribuzione di probabilità di secondo grado, che potrà essere propagata durante l'inferenza per stimare l'incertezza di risultati ottenuti.

Quando si parla di incompletezza dei dati bisogna distinguere tra: informazioni mancanti e informazioni nascoste. Quando sono assenti i valori di una o più variabili si parla di **dati mancanti**. Questi possono essere gestiti in quattro modi (Sucar, 2015):

1. Eliminando i casi con i valori mancanti o inizializzandoli ad uno speciale valore "*unknown*". Entrambe le alternative sono valide se si può contare su una gran quantità di dati poiché altrimenti c'è il rischio di perdere informazioni rilevanti.
2. Sostituire il dato mancante con il valore di moda della variabile; ciò può portare all'introduzione di *bias* nel modello.
3. Stimare il dato mancante basandosi sul valore assunto dalle altre variabili nel *data set*. Questo procedimento rappresenta la miglior alternativa:
 - si inizializza il modello inserendo i dati noti;
 - attraverso il processo d'inferenza si calcola la probabilità a posteriori dei valori mancanti;
 - si assegna ad ogni variabile il valore con la più alta probabilità a posteriori e si costruisce il database completo;
 - si esegue nuovamente l'aggiornamento dei parametri.

Il metodo per gestire i **nodi nascosti** è chiamato *expectation maximization* (EM) (Korb and Nicholson, 2011, pag. 194) e fu introdotto nel 1977 (Dempster et al., 1977). L'assunzione di base è che i dati mancanti o nascosti siano indipendenti dai valori osservati. Questa tecnica prevede la ripetizione di due macro fasi fin tanto che la loro reiterazione porta delle variazioni nella rete: uno *step E - Expectation* -, in cui si stimano i parametri dei nodi non osservabili sulla base dei dati esistenti, ed uno *step M - maximization* -, in cui si aggiornano i parametri tenendo conto dei dati appena stimati. L'algoritmo, nello specifico, è strutturato come segue (Sucar, 2015):

1. Calcolo delle tabelle di probabilità condizionata di tutte le variabili complete con il metodo ML.
2. Inizializzazione dei parametri sconosciuti con valori casuali.
3. Stima del valore dei nodi nascosti svolgendo inferenza sulle variabili note.
4. Aggiornamento dell'intera rete utilizzando i parametri stimati nel punto precedente.

-
5. Calcolo dei valori dei nodi nascosti sulla base dei dati aggiornati.
 6. Il procedimento viene ripetuto fino a che la stima dei parametri continua a variare significativamente.

Questo approccio permette di ottimizzare il calcolo dei valori da attribuire ai nodi non osservati.

2.6.2 Grafo non noto: apprendimento della struttura

Definire ed apprendere la topologia della rete dai dati è un problema complesso e non di semplice risoluzione. Il numero di strutture che possono essere definite è spesso molto elevato e la quantità di dati richiesta per compiere un buon processo di stima dei parametri è sostanziosa e non sempre a disposizione. Gli approcci utilizzabili in questa fase sono due: quello **globale**, che utilizza algoritmi di *search and score* per definire la topologia che nel complesso meglio descrive i dati; quello **locale**, che si basa sullo svolgimento di test d'indipendenza condizionata per valutare le relazioni tra le variabili ed ottenere di conseguenza la struttura che meglio rappresenta queste relazioni. I primi subiscono alterazioni significative quando il numero dei dati a disposizione è basso mentre i secondi sono computazionalmente più complessi.

Nel seguito introdurremo alcune delle tecniche prevalentemente utilizzate per entrambi gli approcci.

Approccio globale: i metodi search and score Gli approcci di tipo globale si basano essenzialmente su tre aspetti: (i) La generazione di differenti strutture adatte a descrivere i dati di partenza. (ii) L'assegnazione un punteggio ad ognuna di esse attraverso la definizione di una funzione di *scoring*. (iii) Il confrontare tra i punteggi delle diverse strutture e la scelta della topologia più rappresentativa. Ipotizzando che le prime due fasi siano state correttamente eseguite, l'ultimo *step* richiede l'utilizzo di un algoritmo di ricerca per determinare la miglior struttura per la rete. Il cuore di queste funzioni è racchiuso nell'implementazione di una ricerca euristica per ridurre il carico computazionale limitando cioè il numero di strutture da valutare. Nella pratica, il numero di possibili strutture di rete è spesso molto elevato, per questo motivo si rende necessario eseguirne una scrematura. Sono di seguito descritti gli algoritmi di ricerca euristica più utilizzati in letteratura.

Hill Climbing Il metodo HC segue un approccio che consiste nel creare una semplice struttura ad albero e migliorarla fino a che non si ottiene la topologia più adatta a descrivere i dati di partenza. Le fasi principali sono (Sucar, 2015):

- (I) Generare una struttura ad albero che costituisca lo scheletro della rete.
- (II) Attribuire un punteggio alla rete iniziale utilizzando una funzione di *scoring*.
- (III) Aggiungere od invertire un arco all'interno della struttura corrente.
- (IV) Calcolare nuovamente il punteggio associato alla rete.
- (V) Eseguire una valutazione: se la misura di adattamento della rete è migliorata, allora si mantiene la modifica apportata; altrimenti si ritorna alla struttura precedente.

(VI) Ripetere il procedimento fino a che si ottengono dei miglioramenti significativi.

Per ottimizzare il processo di apprendimento della struttura alcuni metodi tentano di ridurre il numero di potenziali strutture da valutare tramite l'inserimento di un ordinamento tra i nodi. Il vantaggio risiede nel fatto che gli archi devono seguire le restrizioni imposte dell'ordinamento; ad esempio: se $j > i$ non può esistere un arco che parta dal nodo X_j ed arrivi in X_i . Dato un ordinamento con queste caratteristiche, apprendere la struttura è un'operazione equivalente a ricercare il miglior insieme di genitori per ogni nodo.

K2 L'algoritmo K2 sfrutta questo tipo di ordinamento per ottenere un vantaggio in termini di efficienza computazionale e per garantire l'assenza di cicli nel modello.

Brevemente⁶ l'approccio consiste nel ricevere in *input* il database D con N osservazioni sull'insieme di variabili X_1, X_2, \dots, X_p ordinate e il numero massimo di genitori per ogni nodo, u . All'inizio, tutti i nodi sono inizializzati come *root* quindi sono sprovvisti di genitori. Partendo dalla prima variabile e seguendo l'ordinamento, per ogni nodo l'algoritmo testa tutti i possibili genitori che non sono stati ancora aggiunti e vi assegna quello che massimizza il punteggio globale. Il procedimento è reiterato fino a che non finisco i genitori possibili che apportano maggior valore alla rete. Il *set* di genitori, $Pa(X_i)$, per ogni nodo X_i del modello, rappresenta l'*output* e definisce la struttura della rete. Per misurare il grado di adattamento della topologia di una rete ai dati iniziali sono state ideate le funzioni di SCORING. Le caratteristiche ideali per una funzione di questo tipo sono (Sucar, 2015):

- + Scomponibilità: il punteggio finale assegnato alla rete è calcolato come somma dei singoli valori locali che dipendono solamente da ogni nodo e dai suoi genitori.
- + Punteggi equivalenti: a DAG che descrivono essenzialmente lo stesso grafo, cioè quando rappresentano le stesse relazioni di indipendenza, è assegnato lo stesso punteggio finale.

Di seguito sono brevemente descritte le funzioni di *scoring* più utilizzate in letteratura⁷.

ML - Maximum likelihood Questa funzione (Korb and Nicholson, 2011, pag. 195) rappresenta la probabilità di un insieme di dati D data la struttura G :

$$ScoreML = P(D | \Theta_G, G_i) \quad (2.19)$$

dove G_i indica la struttura candidata e Θ_G il corrispondente vettore di parametri⁸. L'*output* prodotto rappresenta il set di parametri Θ che massimizza la *maximum likelihood* della rete G . La tendenza a privilegiare le reti più complesse, poiché descrivono meglio il set di dati D , è lo svantaggio maggiore di questo metodo.

⁶Per approfondimenti si veda: (Sucar, 2015)

⁷Per maggiori approfondimenti si veda (De Campos, 2006), (Sucar, 2015)

⁸Il vettore contiene le tabelle di probabilità condizionata di ogni nodo associate alla i-esima struttura G_i .

BIC - Bayesian information criterion Nel tentativo di trovare un compromesso tra precisione e complessità del modello è nata la funzione BIC, definita dalla formula:

$$BIC = \log P(D | \Theta_G, G_i) - \frac{d}{2} \log N \quad (2.20)$$

dove d è il numero di parametri nelle BN e N indica la numerosità campionaria nel data set. Al contrario della ML questa funzione integra una forma di penalizzazione per le reti più complesse con il rischio, però, di selezionare topologie troppo semplici. Si noti che il BIC non richiede in *input* le probabilità a priori associate alle variabili.

AIC - Akaike information criterion Questa funzione è simile alla precedente, la differenza risiede nell'adozione di un fattore di penalizzazione per la complessità più semplice:

$$AIC = \log P(D | \Theta_G, G_i) - d \quad (2.21)$$

MDL - Minimum description length Il metodo del MDL, definito da Rissanen (Rissanen, 1978), è equivalente al BIC.

BS - Bayesian score Appartengono a questa categoria tutte le funzioni che seguono l'approccio Bayesiano ottenendo, attraverso il teorema di Bayes, la probabilità a posteriori della topologia partendo dall'insieme dei dati:

$$P(G_i | D) = P(G_i)P(D | G_i)/P(D) \quad (2.22)$$

dove $P(G_i)$ è la probabilità a posteriori della struttura G_i e può essere inizializzata ad una distribuzione uniforme oppure definita da un insieme di esperti.

BDE - Bayesian Dirichlet equivalent score Il BDE è una variante dei metodi Bayesiani che si basa su delle assunzioni specifiche:

1. I parametri sono indipendenti e si ipotizza che a priori seguano la distribuzione di Dirichlet.
2. La funzione attribuisce punteggi equivalenti.
3. I dati campionari sono indipendenti e identicamente distribuiti.

Considerazione una generica configurazione in cui $X_i = k$ dato $Pa(X_i) = j$ e N' rappresenta la dimensione del campione, il punteggio ad essa attribuito è calcolato con la seguente formula:

$$N_{ijk} = P(X_i = k, Pa(X_i) = j | G_i, \Theta_G) x N' \quad (2.23)$$

K2 score La funzione di *scoring* K2⁹ è utilizzata nell'omonimo algoritmo di ricerca. Rappresenta una semplificazione delle BS e consiste nello scomponere il punteggio finale calcolandolo per ogni variabile X_i noto il set dei suoi genitori, $Pa(X_i)$, ottenendo:

$$S_i = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} \alpha_{ijk}! \quad (2.24)$$

⁹Cooper and Herskovits (1992)

dove r_i è il numero di valori che può assumere X_i , q_i è il numero totale di possibili combinazioni dei nodi genitori di X_i , α_{ijk} è il numero di casi nel database in cui $X_i = K$ quando $Pa(X_i) = j$, infine N_{ij} è la frequenza con cui $Pa(X_i) = j$.

Approccio locale: i metodi constraint-based I metodi *constraint-based* si basano sull'apprendimento della struttura analizzando le relazioni probabilistiche che intercorrono tra le variabili. L'analisi è svolta attraverso dei test di indipendenza condizionata. Il risultato finale è un grafo che rappresenta le relazioni emerse dai test.

Di seguito sono descritti gli algoritmi più utilizzati.

PC Algorithm In un primo momento, il PC costruisce lo scheletro della rete dai dati e poi si focalizza sul direzionare gli archi. Il procedimento è riassumibile in tre fasi:

1. Definizione di un grafo non orientato che costituisce lo 'scheletro' della rete. Si svolgono dei test per capire se due variabili, X ed Y , sono condizionatamente indipendenti dato uno specifico sottoinsieme S di nodi, $I(X, Y | S)$. I test devono essere svolti per ogni coppia di nodi nel modello.
2. Fissare la direzione degli archi sulla base dei test di indipendenza condizionata eseguiti su terzine di variabili. L'algoritmo ricerca nel grafo sotto-strutture del tipo $X - Z - Y$, in cui non esiste un arco tra X ed Y . A questo punto, se X ed Y non sono indipendenti dato Z , $(X \not\perp\!\!\!\perp Y | Z)$, significa che fanno parte di una *v-structure* del tipo $X \rightarrow Z \leftarrow Y$.
3. Impostare le rimanenti frecce rispettando l'obbligo di aciclicità e i vincoli imposti dai risultati dei precedenti test.

Il risultato finale è un grafo fedele ai dati iniziali, dai quali è stato generato. L'efficacia di questo metodo è legata alla quantità di informazioni presenti nel *database*: maggiore è la numerosità campionaria e più alta è la precisione dei test.

2.6.3 Combinazione di approcci: conoscenze degli esperti e dati

La procedura di apprendimento della topologia è fortemente condizionata dalla quantità di dati a disposizione. Quando le informazioni iniziali non sono sufficienti è necessario integrarle con dati provenienti da altre fonti; le conoscenze dei così detti "esperti", ad esempio, sono largamente utilizzate nella costruzione delle BN.

L'utilità di combinare le informazioni in fase di apprendimento dei parametri è evidente ogni qual volta i parametri siano incerti o siano parzialmente assenti¹⁰. Per quanto riguarda l'apprendimento della struttura, invece, le conoscenze degli esperti possono essere integrate seguendo due approcci:

- **Restrittivo:** le informazioni svolgono una funzione di restrizione riducendo il numero di topologie da analizzare. In questo modo gli algoritmi risulteranno più efficaci ed efficienti. Per esempio, possono esser usate per: (i) Determinare l'ordinamento causale delle variabili. (ii) Vincolare le direzioni di specifici archi. (iii) Indicare se esistano archi che possano essere direzionati in entrambi i sensi. (iv) Evitare che due variabili siano direttamente collegate. (v) Qualsiasi combinazione dei precedenti casi.

¹⁰Si veda la sezione sull'apprendimento dei parametri di questo capitolo.

-
- **Propositivo:** i dati raccolti sono utilizzati per validare una struttura proposta dagli esperti.

2.7 Estensioni del modello

Il modello di rete Bayesiana descritto finora è adatto a rappresentare le relazioni intercorrenti tra gli elementi del sistema osservato in un istante di tempo preciso. Nonostante le grandi potenzialità di questo strumento, esso presenta limiti dal punto di vista funzionale. Si potrebbe per esempio essere interessati ad utilizzare una BN come strumento di supporto in un processo decisionale o a integrare dati raccolti in periodi diversi.

La versatilità di cui godono le reti Bayesiane ha permesso d'implementare estensioni del modello di base per soddisfare anche altri ordini di esigenze.

2.7.1 Diagrammi di influenza

Le reti Bayesiane, come ogni altro modello probabilistico, nascono per rappresentare la realtà in condizioni d'incertezza. In un processo decisionale sono sempre presenti dubbi o ambiguità che portano distorsioni, ragion per cui sono nati gli *Influence Diagram (ID)*, una particolare tipologia di rete decisionale.

A differenza delle BN, essi integrano sia le azioni da valutare sia le utilità attese come risultato di ogni azione. Attraverso un algoritmo *ad hoc* è possibile introdurre le evidenze nel modello e inferenza per stimare l'insieme delle decisioni che probabilmente è conveniente intraprendere.

Prima di proseguire, si vuole fornire una definizione formale di diagramma d'influenza o *Influence Diagram (ID)* (Kjaerulff and Madsen, 2008).

Definizione 2.23 (Influence diagram). *Un diagramma d'influenza $N = (X, G, P, U)$ è composto da:*

- *un DAG, $G = (N, A)$ con un set di nodi, N , ed un set di archi, A , indicante le relazioni causali, le informazioni note prima di prendere una decisione (archi d'informazione) e l'ordine sequenziale delle scelte (archi di precedenza).*
- *l'insieme dei nodi X_C casuali e degli X_D nodi decisionali in modo tale che $X = X_C \cup X_D$ rappresenti tutti i nodi della rete.*
- *il set P con le distribuzioni di probabilità condizionata per ogni variabile casuale, $P(X_i | Pa(X_i))$.*
- *l'insieme delle utilità U , contenenti una funzione di utilità, $u(Pa(X_i))$, associata ad ogni nodo facente parte del sotto insieme $N_U \subset N$ dei nodi utilità.*

Utilità attesa Alla base di un ID è posta l'ipotesi di razionalità degli individui; ciò significa che il loro comportamento è focalizzato sulla massimizzazione dei benefici o sulla minimizzazione dei costi. Questo ragionamento è valido solo perché esistono le **funzioni di utilità** che associano un numero reale al risultato di ogni azione calcolato in base alle conseguenze che ne derivano (Sucar, 2015). Basandosi sulle preferenze dei singoli soggetti, le funzioni sono diverse per ogni individuo, visto che soggetti diversi attribuiscono ad una

stessa decisione un livello di utilità diverso. La teoria delle probabilità combinata con la mappatura delle utilità definisce il concetto di utilità attesa, utilizzato per selezionare la decisione migliore.

Definizione 2.24 (Utilità attesa). *Data l'evidenza E , l'utilità attesa di A , azione non deterministica i cui possibili risultati sono indicati con O_i , è calcolato come:*

$$EU(A | E) = \sum_i P(O_i | E, A)U(O_i | A) \quad (2.25)$$

dove: $U(O_i | A)$ è l'utilità associata ad ogni conseguenza sapendo che l'azione A è stata intrapresa mentre $P(O_i | E, A)$ rappresenta la distribuzione di probabilità condizionata dei possibili risultati di A sapendo che è stata osservata l'evidenza E e che l'azione A è stata compiuta.

Tenendo a mente l'ipotesi iniziale sulla razionalità dei soggetti, ciascun individuo selezionerà il set di azioni che massimizzano l'utilità attesa. Questo è il **principio di massimizzazione dell'utilità attesa**.

Concetti base delle reti decisionali Le reti decisionali, come estensioni delle reti Bayesiane, rappresentano sistemi in cui si richieda d'intraprendere delle decisioni. Per raggiungere questo scopo, sono introdotte nuove tipologie di nodi (Figura 2.7):

- + Opportunità (*Chance node, X_C*): rappresentati da una forma ovale, indicano le variabili casuali esattamente come in una BN qualunque, ciascuna con la propria tabella di probabilità condizionata. Possono essere figli sia di nodi opportunità sia di nodi decisionali.
- + Decisioni (*Decision node, X_D*): indicano le scelte da prendere in un determinato istante e graficamente hanno una forma rettangolare. I valori che possono assumere sono le azioni che possono essere intraprese. Se la rete contiene un solo nodo decisionale, allora avrà come genitori dei *chance node*. Nel caso ci siano più scelte, queste possono essere figlie di altri nodi decisionali, per indicare l'ordine con cui devono essere compiute.
- + Utilità (*Utility node, X_U*): rappresentati da un diamante, esprimono l'utilità o il costo di ogni decisione. I genitori di un nodo di utilità sono le variabili che direttamente influiscono sul suo valore. Ad ognuno di questi nodi è associata una tabella di utilità, la quale contiene il valore assegnato ad ogni possibile combinazione degli stati dei suoi genitori. Questo tipo di variabile non può avere figli.

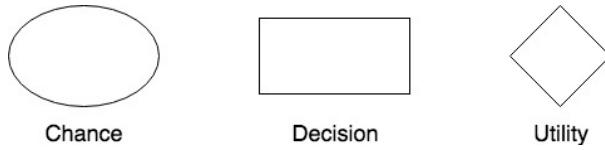


Figura 2.7: Tipologie di nodi nelle reti decisionali

In un ID non tutti gli archi hanno lo stesso significato. Gli **archi di precedenza** (*precedence arc*) collegano due nodi decisionali ed esprimono l'ordine sequenziale in cui compiere

le scelte. Un **arco informativo** (*information arc*) termina in un nodo decisionale e delimita il set dei suoi antenati con cui si esprimono le informazioni note prima di compiere la scelta. Con questi archi si crea la **tabella decisionale** contenente l'azione ottima da intraprendere per ogni combinazione dei suoi antenati. Da un ID si può derivare la corrispondente BN rimuovendo gli archi d'informazione.

I valori di un nodo decisionale sono o di tipo **non interveniente**, quindi non influenzare direttamente un nodo opportunità, o **interveniente**, che hanno conseguenze dirette su contesto analizzato.

La regola del prodotto in un diagramma d'influenza è composta da:

$$EU(X) = \prod_{X_v \in X_c} P(X_v | Pa(v)) \sum_{w \in V_u} u(Pa(w)) \quad (2.26)$$

Il processo di valutazione delle reti con un singolo¹¹ *decision node* comincia con l'introduzione delle evidenze disponibili. Per ogni azione A_i , contenuta nel nodo decisionale si applica un algoritmo d'inferenza per calcolare la probabilità a posteriori dei genitori del nodo utilità. Infine, tramite l'Equazione 2.25 si determina l'utilità attesa associata alla specifica azione A_i . Il procedimento si conclude restituendo in *output* l'azione con l'utilità attesa più elevata.

Se il processo decisionale richiede di compiere più scelte, allora bisogna ricorrere a metodi più sofisticati per determinare l'insieme delle azioni da intraprendere. Per esempio si possono usare: *modelli di combinazione test-azioni* (Korb and Nicholson, 2011, pag. 106), *reti decisionali dinamiche* (Korb and Nicholson, 2011, pag.118), ecc.

2.7.2 Dynamic Bayesian Networks (DBNs)

Una BN è una rappresentazione istantanea delle relazioni causali in un sistema. Sostanzialmente compiono una fotografia del fenomeno, escludendo le relazioni temporali, quindi che il valore di un nodo all'istante t possa influenzare lo stato di una o più variabili in un momento $t+1$ successivo. Le reti Bayesiane dinamiche¹² (*dynamic Bayesian network (DBN)*) monitorano i cambiamenti che avvengono nel sistema in un periodo di tempo.

Supponendo che il dominio analizzato sia composto da n variabili casuali $X = X_1, X_2, \dots, X_n$, la costruzione di una DBN prevede che sia aggiunto un nodo, indicato con X_i^t , per ogni variabile X_i e per ogni *step* temporale t considerato. In un modello dinamico si suppone che lo stato del sistema in t influirà sullo stato futuro delle variabili, in $t+1$ ma dipenderà a sua volta dalla passata configurazione in $t-1$.

Il tempo è diviso in ***time-slice***, cioè in periodi di tempo. La relazione tra due variabili collegate nello stesso *time-slice* si dice ***intra-slice***, $X_i^t \rightarrow X_k^t$.

I legami tra nodi appartenenti a due *time-slices* diversi sono evidenziate da archi ***inter-slice*** (o archi temporali) e includono le relazioni sia della stessa variabile nel tempo ($X_i^t \rightarrow X_i^{t+1}$) sia di due variabili distinte ($X_i^t \rightarrow X_k^{t+1}$). Lo stato di un qualsiasi nodo in un istante può condizionare il valore di una o più variabili in un periodo successivo ma mai precedente; ciò significa che può impattare solo sullo stato del sistema futuro.

Gli archi temporali non possono estendersi per più di un *time-slice* perciò collegano solo nodi appartenenti a periodi adiacenti. Questo vincolo strutturale, derivante dal rispetto della proprietà di Markov, stabilisce che i valori assunti dai nodi della rete al tempo t

¹¹Per la risoluzione di reti più complesse si veda: (Korb and Nicholson, 2011)

¹²Per maggiori dettagli si veda: (Kjaerulff, 1995)

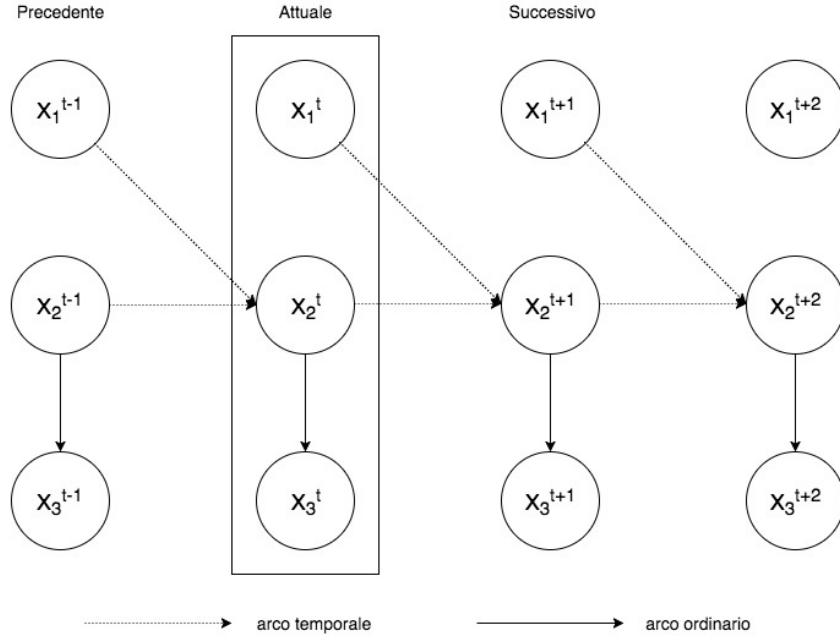


Figura 2.8: Esempio di DBN

dipendono solo ed esclusivamente dalle scelte compiute in t e dallo stato assunto in precedenza dal sistema.

La parte quantitativa del modello è formata dalle CPT di ogni nodo che sono calcolate come segue:

$$P(X_i^t | Y_1^t, \dots, Y_m^t, X_i^{t-1}, Z_1^{t-1}, \dots, Z_r^{t-1}) \quad (2.27)$$

dove Y_1^t, \dots, Y_m^t sono i genitori *intra-slice* di X_i^t , $Z_1^{t-1}, \dots, Z_r^{t-1}$ sono i genitori *inter-slice* di X_i^t e X_i^{t-1} indica lo stato assunto dalla variabile nel *time slice* precedente.

Per compiere qualsiasi tipo di ragionamento inferenziale bisogna utilizzare degli algoritmi specifici per la propagazione delle evidenze in questi modelli.

Dynamic decision network (DDN) Un cenno¹³ ad un’ulteriore estensione delle reti Bayesiane riguarda le reti decisionali dinamiche (*dynamic decision network*, DDN). Esse modellano i cambiamenti che avvengono in un contesto quando sono chiamati in gioco processi decisionali che implicano delle scelte sequenziali.

L’obiettivo è determinare la successione di scelte ottime per massimizzare l’utilità nell’ultimo *time slice*.

2.7.3 Object Oriented Bayesian Network (OOBNs)

Analizzando problemi complessi in cui il numero di variabili è elevato e sono chiamate in gioco enormi quantità di dati, le reti Bayesiane classiche possono risultare inefficienti. Per superare quest’ostacolo, durante la costruzione del modello si può optare per l’implementazione di una *object oriented Bayesian network* (OOBN), una rete “orientata ad oggetti”. A differenza di una semplice BN, una OOBN è formata da **classi** definite (Kjaerulff and Madsen, 2008, pag.92) come:

¹³Per approfondimenti si veda (Korb and Nicholson, 2011, pag.118)

Definizione 2.25 (OOBN class). In una OOBN una generica classe $C = (GIHO)$ consiste in:

- una rete probabilistica che collega le X variabili attraverso un DAG G .
- un set di variabili in input tali che $I \subseteq X$.
- dei nodi di output in modo che $I \cap O = 0$.
- i nodi che non appartengono né ad I né ad O costituiscono l'insieme dei nodi nascosti tali che $H = X \setminus (I \cup O)$.

Ogni classe può contenere sia nodi ordinari sia oggetti, cioè istanze di altre classi. Un **oggetto** può includere altri oggetti al suo interno dando forma ad una struttura gerarchica che consente di semplificare le fasi di elaborazione e modellazione del problema. Sono chiamati **nodi di interfaccia** le variabili di un oggetto che lo collegano agli altri elementi della rete. Nel caso in cui questi nodi costituiscano un sotto insieme dei nodi della classe, si dice che l'informazione è nascosta poiché il resto delle variabili non è visibile al di fuori della classe. Gli *interface node* sono divisibili in due categorie: quelli di **input**, che sono i nodi radice di una determinata classe e che devono essere mappati pari pari ai corrispettivi nodi all'esterno della classe a cui sono collegati; quelli di **output**, cioè quelle variabili che diventeranno genitori dei nodi all'esterno della classe.

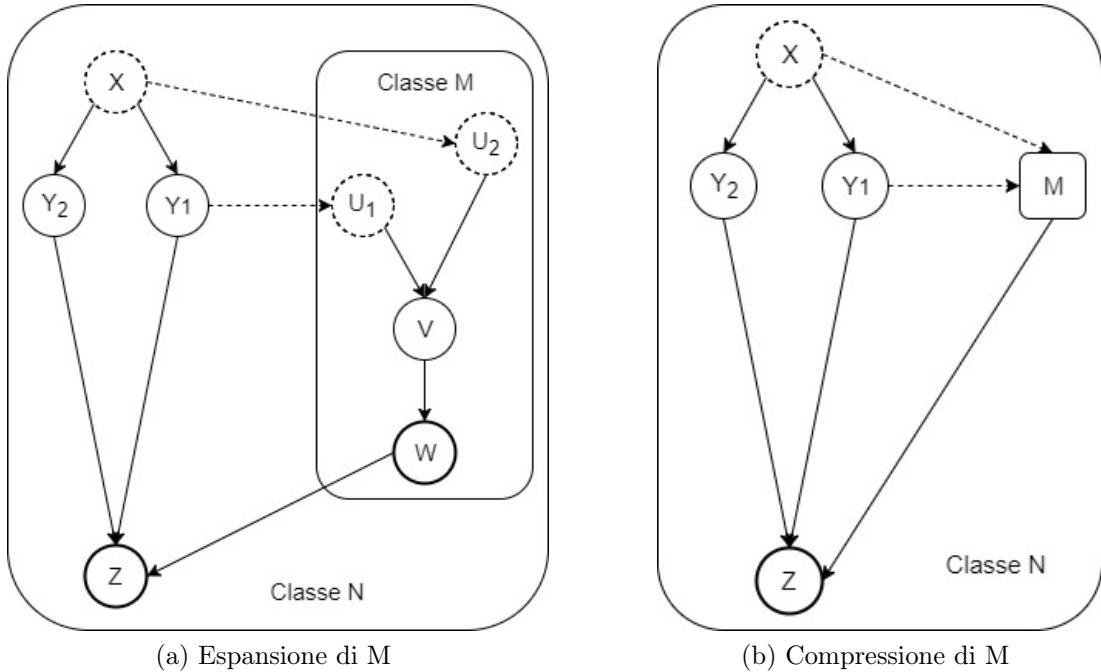


Figura 2.9: Esempio di OOBN

La Figura 2.9 mostra una semplice OOBN in cui sono indicati: gli oggetti con un rettangolo, i nodi di *input* con un cerchio tratteggiato, i nodi *output* con un cerchio con il bordo spesso e i nodi nascosti con dei cerchi ordinari.

Una rete Bayesiana orientata ad oggetti è riconducibile ad una classica BN attraverso un procedimento ricorsivo in cui si sostituisce ogni *object* con la corrispondente struttura di nodi ed archi sottostante. A questo punto, l'aggiornamento dei parametri è compiuto attraverso un qualsiasi algoritmo inferenziale.

2.8 Vantaggi

Le reti Bayesiane sono largamente utilizzate in ambiti anche molto diversi tra loro. Questa è la dimostrazione che sono uno strumento inferenziale e decisionale estremamente flessibile. Esse permettono di (Cenatiempo et al., 2010):

- integrare informazioni provenienti da fonti diverse (conoscenze teoriche, opinioni di esperti, dati sperimentali, ecc.).
- raggiungere una maggior efficienza computazionale, poiché richiedono di inserire le probabilità condizionate delle sole variabili legate tra loro da un arco.
- eseguire operazioni che vanno al di là della capacità della mente umana, come l'aggiornamento dei parametri della rete a seguito della sopravvenuta conoscenza di nuove informazioni.
- stabilire e controllare le assunzioni alla base della rete e del sistema per limitare la validazione delle conclusioni.
- compiere analisi di sensibilità utili per quantificare l'influenza di una variabile sull'intera rete.

Il più grande limite, che per lungo tempo ha afflitto questo strumento, era l'elevato carico computazionale richiesto in alcune elaborazioni; ad esempio, quando sono utilizzate delle variabili casuali continue. Si può affermare che quest'ostacolo è stato ormai superato per mezzo sia delle moderne tecnologie sia delle tecniche e degli algoritmi di elaborazione sempre più efficienti.

Capitolo 3

Le reti Bayesiane per l'analisi della soddisfazione dei cittadini

Dal 2001 il comune di XXX¹ si è impegnato per migliorare il proprio servizio di raccolta e smaltimento dei rifiuti. In particolare, l'attivazione di nuove modalità di conferimento e di raccolta dei rifiuti solidi urbani ha portato ad una maggiore sensibilizzazione dei cittadini. Segnali positivi ed incoraggianti sono, ad esempio, l'aumento significativo della percentuale di materiali riciclati ogni anno.

L'amministrazione comunale è convinta che, dopo tre anni, questo servizio sia suscettibile di ulteriori miglioramenti. Nel 2004, la sezione 'Direzione Lavori Pubblici' dell'amministrazione comunale di XXX ha affidato al proprio ufficio competente, l'Ufficio 'Ecologia e Verde Pubblico', il compito di verificare, attraverso lo svolgimento di un'indagine *ad hoc*, il grado di soddisfazione dei propri cittadini sul servizio di raccolta dei rifiuti.

Lo scopo di quest'analisi consiste nel valutare l'efficacia delle BN nel modellare i dati raccolti dall'indagine al fine di creare scenari utili per le possibili politiche da adottare per migliorare il grado di soddisfazione dei cittadini.

3.1 Raccolta dei dati

La raccolta delle informazioni è stata affidata ad un'azienda esterna che ha svolto un'indagine CATI² sottponendo agli intervistati un questionario preparato appositamente. Il soggetto esterno addetto alla raccolta dei dati ha garantito che il campione estratto è rappresentativo della popolazione comunale. Per ogni nucleo familiare coinvolto nell'indagine è stato chiesto che le risposte fossero date da colui che si occupa regolarmente della gestione dei rifiuti domestici.

Il questionario è suddiviso in sette parti, ognuna delle quali sarà dettagliatamente analizzata in seguito:

1. Dati anagrafici dell'intervistato;
2. Informazione sul servizio, per valutare il grado d'informazione dei cittadini;
3. Ruolo degli operatori, per monitorare il loro operato;

¹Il comune non è indicato per motivi di riservatezza.

²CATI: modalità di rilevazione diretta di unità statistiche realizzata attraverso interviste telefoniche, dove l'intervistatore legge le domande all'intervistato e registra le risposte su un computer, tramite un apposito software.

-
4. Valutazione del servizio, per raccogliere i dati sull'importanza e la soddisfazione dei fattori caratterizzanti il servizio;
 5. Conoscenza ed utilizzo degli altri servizi accessori offerti dal comune, oltre a quello della raccolta dei rifiuti;
 6. Livello di contribuzione, per valutare l'incidenza di una variazione delle tasse sulla soddisfazione dei cittadini;
 7. Servizio di raccolta rifiuti nel centro storico.

Il questionario usato nell'indagine è riportato in Appendice B.

La scelta di utilizzare un questionario per la raccolta dei dati presenta il vantaggio che un sondaggio composto da n domande produce delle risposte che possono essere riassunte all'interno di altrettante variabili casuali X_1, X_2, \dots, X_n . A loro volta, le variabili possono essere divise in due gruppi: da un lato le q variabili *target*, contenenti informazioni generali riguardanti la soddisfazione complessiva, dall'altra le restanti $n - q$ variabili che sono analizzate ricercando una correlazione con le *target*. Bisogna verificare, attraverso l'impiego di metodi statistici e di *data mining*, se le combinazioni (X_i, X_j) con $X_i \in X_1, \dots, X_{n-q}$ e $X_j \in X_{n-q+1}, \dots, X_n$ sono dipendenti o indipendenti per ogni coppia di variabili (Hand et al., 2001).

3.2 Tipologia dei dati

I risultati della fase di raccolta dei dati sono archiviati all'interno di un *database* composto da 77 variabili (colonne) e 509 osservazioni (righe) ciascuna delle quali individua le risposte date da un singolo intervistato.

3.2.1 Operazioni preliminari sui dati

Durante l'analisi di un fenomeno, uno degli errori più comuni è dare per scontato che i **dati grezzi** (*raw data*) raccolti attraverso l'indagine siano corretti, dove per "corretti" s'intende: privi di errori, registrati in un formato corretto e completi. Quest'ipotesi difficilmente è verificata quindi prima di compiere qualsiasi analisi di tipo statistico è necessario investire del tempo per preparare i dati.

La **pulizia dei dati** (*Data Cleaning*) è il processo che trasforma i dati grezzi in dati consistenti che possono essere sottoposti ad analisi statistiche. L'obiettivo è migliorare il contenuto e l'attendibilità delle conclusioni basate sull'analisi dei dati attraverso l'eliminazione delle inconsistenze presenti nella banca dati. Le operazioni più frequenti riguardano l'imputazione dei dati mancati (*NA*) e la gestione dei dati anomali (*outliers*). Il problema principale di questo tipo di operazioni è che possono influenzare significativamente i risultati delle elaborazioni; bisogna prestare attenzione ad evitare di alterare eccessivamente il significato intrinseco nei dati iniziali (De Jonge and Van Der Loo, 2013). A partire dai dati grezzi ottenuti come risultato del sondaggio, il processo si suddivide generalmente in due fasi:

1. **Dati tecnicamente corretti:** trasformazione dei dati grezzi in *technically correct data* correggendo eventuali intestazioni di colonne mancanti, inserimenti di dati in variabili di tipo non coerente (es: inserire l'età in una variabile di caratteri

invece che in una numerica), registrazione di etichette categoriali non previste o non conosciute, utilizzo di codifiche errate.

2. **Dati consistenti:** non essendoci alcuna garanzia che i valori siano completi e privi di errori non si possono ancora eseguire analisi statistiche fino a che non si ottengono dei *consistent data*. Al termine, saranno pronti per essere elaborati e produrre dei "risultati statistici". Le inconsistenze sono generate dalla presenza di valori mancanti, dati anomali, errori o valori speciali che devono essere innanzitutto identificati e poi corretti.

L'articolazione del processo appena descritto a grandi linee, varia caso per caso a seconda delle caratteristiche del *database* da analizzare.

Nello specifico della soddisfazione del servizio di raccolta dei rifiuti, il primo passo è analizzare ogni variabile con lo scopo di verificare la presenza di errori tecnici nei dati grezzi. Tutte le variabili sono di tipo categoriale o discreto quindi l'attenzione è posta sul controllo delle etichette in modo che non siano registrate delle *labels* sconosciute o errate. Successivamente si identificano le variabili contenenti informazioni superflue o ridondanti, considerando ridondanti tutte le colonne già registrate in altre variabili oppure risultanti da un processo di calcolo deterministico. Questo ragionamento ha permesso di escludere otto variabili come ad esempio la nazionalità degli intervistati, avendo rilevato che il 98% di essi sono italiani. La banca dati è stata ridotta quindi a 509 osservazioni e 69 variabili. I **dati mancanti** o *Not available (NA)* rappresentano elementi di natura nota ma dei quali non è stato possibile reperire il valore. Nella prassi sono gestiti utilizzando, singolarmente o in combinazione, i seguenti metodi (De Jonge and Van Der Loo, 2013):

- Eliminazione: quando il loro numero all'interno di una riga o colonna è elevato si può optare per eliminare interamente l'osservazione o la variabile dato il suo basso apporto informativo. In questi casi bisogna verificare che il numero di valori mancanti si distribuisca casualmente e che non si sta invece eliminando una categoria di individui.
- Ripristino del valore: quando il contesto lo consente, bisogna cercare di recuperare il dato mancante, ad esempio richiamando l'intervistato come possibile nel caso di un sondaggio CATI;
- Imputazione: consiste nel sostituire il dato mancante con un valore sostituto scelto attraverso uno dei criteri esistenti. Quelli più frequentemente utilizzati nella pratica:
 - Media: il valore che sostituisce il dato mancante è la media;
 - Punto comune: come sostituto si utilizza un valore come la moda o la mediana;
 - Regressione: il dato inserito al posto di quello mancante è la previsione ottenuta come risultato del processo di regressione semplice o multipla;
 - Algoritmo di *data mining*: è l'approccio più sofisticato e prevede l'utilizzo di algoritmi, come il *k nearest neighbours (kNN)*, per determinare quale sia il valore sostituto che meglio si adatta ad essere inserito in quella posizione.

Come già anticipato, la scelta del metodo (o dei metodi) con cui si gestiscono i dati mancanti influenza in modo più o meno significativo i risultati dell'indagine; per questo motivo occorre compiere questa scelta con attenzione per evitare l'alterazione dei dati e la conseguente perdita di attendibilità delle conclusioni.

Per cominciare, analizzando la distribuzione degli NA nelle varie sezioni del questionario, si nota che essa non subisce significative variazioni all'interno di 'segnalazioni, reclami e suggerimenti' e 'raccolta nel centro storico'. Da ulteriori approfondimenti è emerso che questa comportamento è dovuto alla presenza di particolari categorie d'individui che sono, rispettivamente, coloro che non hanno mai compiuto alcuna segnalazione e coloro che non risiedono nel centro storico o che anche se vi risiedono non considerano la raccolta porta a porta dei rifiuti nel centro storico una delle cause di perdita di decoro. Queste informazioni non rappresentano propriamente degli NA ma delle caratteristiche intrinseche di due gruppi di intervistati: per questo motivo i dati mancanti sono sostituiti con un valore che identificasse questa particolarità.

Per applicare il metodo dell'eliminazione bisogna prima definire la soglia oltre cui l'apporto informativo di un'osservazione è considerato basso e quindi trascurabile. Il limite impostato è 20% del numero di variabili: $Soglia = numvariabili * 20\% = 69 * 0.2 = 14$. Dopo aver calcolato rapidamente il numero di dati mancanti contenuti in ogni riga, sono eliminate le osservazioni che non rispettavano questa soglia, nove in totale. Successivamente sono eliminate altre osservazioni riducendo la dimensione del *database* a 480 righe, risultato accettabile che non pregiudica le conclusioni visto che l'eliminazione non ha coinvolto più del 5% delle osservazioni.

Gli errori più comuni commessi dagli analisti in questa fase sono due: il primo è quello di scambiare il dato NA per una categoria o un valore di *default*; il secondo è confondere, nelle variabili categoriali, il dato mancante con la categoria "sconosciuto" (*unknown*). Per non incorrere in questo secondo errore, sono stati individuati i due casi in cui i dati mancati sono equivalenti a risposte "non so, non risponde": il primo relativo alla professionalità degli operatori e il secondo riguardante il decoro del centro storico.

A questo punto il processo di pulizia dei dati è concluso e i dati possono essere elaborati. Per quanto concerne la tipologia delle singole variabili, esse sono classificate come segue:

- Numeriche discrete: le valutazioni sull'importanza e la soddisfazione dei fattori e degli aspetti sono espresse in termini numerici, quindi le corrispondenti variabili sono considerate discrete e numeriche;
- Categoriale: le risposte a tutte le restanti domande del questionario sono contraddistinte da etichette, motivo per cui ricadono all'interno delle variabili categoriali.

L'importanza di questa distinzione sarà chiara nel proseguo della tesi nella sezione dedicata all'analisi delle correlazioni o all'apprendimento della struttura della rete. Per inquadrare meglio i dati si procede con l'analisi di ciascuna variabile.

3.2.2 Dati anagrafici

Le informazioni raccolte in questa sezione permettono di identificare e classificare il campione a cui è sottoposto il questionario.

Genere Delle 480 persone intervistate, 127 sono uomini mentre 353 sono donne (Figura 3.1). Siccome le risposte sono date da chi si occupa dello smaltimento dei rifiuti domestici si può concludere che nel 73.54% dei casi sono le donne che prevalentemente si occupano di svolgere quest'attività.

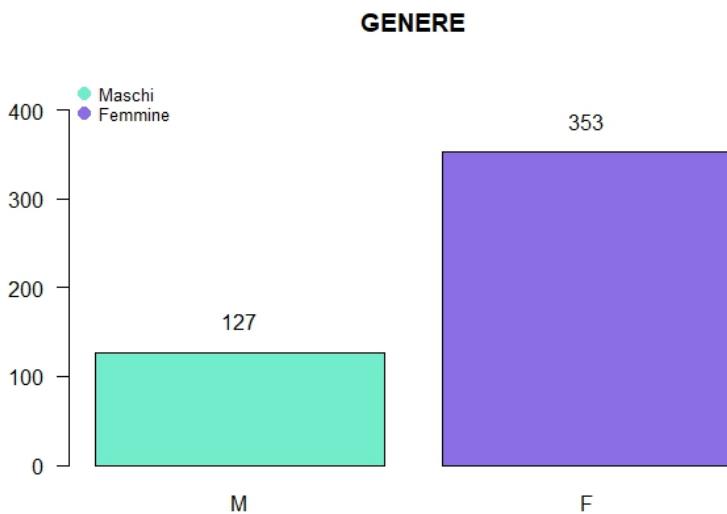


Figura 3.1: Genere

Classe età	Freq. Assoluta
< 25	7
26 - 35	21
36 - 45	124
46 - 55	202
> 56	126

Tabella 3.1: Classi d'età

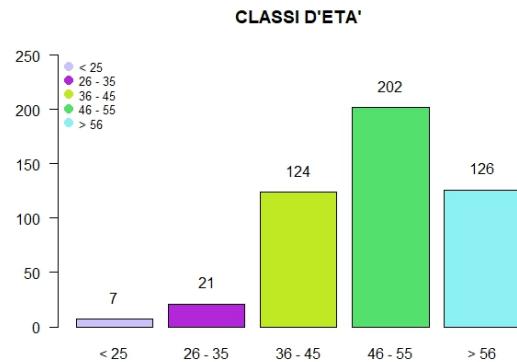


Tabella 3.2: Frequenza distribuzione per classi d'età

Classe d'età Sulla base dell'età, i cittadini sono suddivisi in cinque fasce e i risultati sono riportati nella Tabella 3.1. La classe modale³ comprende coloro che hanno un'età compresa tra i 46 ed i 55 anni con una frequenza assoluta di 202 intervistati. La distribuzione è asimmetrica, la maggior parte dei dati è collocata nella parte superiore della distribuzione. Si può concludere che i giovani sono coloro che per ultimi si occupano dello smaltimento dei rifiuti all'interno delle mura domestiche.

Titolo di studi Un ulteriore elemento di discriminazione è il titolo di studi raggiunto. Il conseguimento di un titolo implica l'aver ottenuto anche quelli precedenti; ad esempio: chi si è laureato, necessariamente ha terminato con successo la scuola dell'obbligo e quella superiore. Le frequenze assolute sono riassunte nella Figura 3.2. Il 50.83% del campione ha terminato con successo la scuola dell'obbligo mentre solo il 12.5% ha conseguito una laurea. Ricordando che la media dell'età degli intervistati è alta, la distribuzione di questa variabile è probabilmente legata al fatto che, rispetto ad oggi, una volta il numero di coloro che decidevano di intraprendere un percorso di laurea era basso.

³Classe modale: se la distribuzione della variabile è suddivisa in classi, allora la classe modale è definita come la classe alla quale corrisponde la frequenza più alta (Borra and Di Ciaccio, 2008).

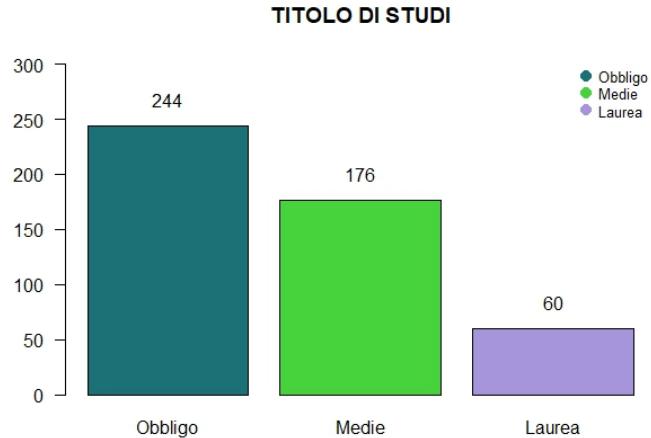


Figura 3.2: Titolo di studi

Attività lavorativa Una delle domande del questionario è finalizzata a conoscere l’impiego professionale dell’intervistato, distinguendo tra: (a) Lavoratore dipendente; (b) Lavoratore autonomo; (c) Disoccupato; (d) Pensionato; (e) Studente; (f) Casalinga. I risultati ottenuti sono contenuti nel grafico a barre riportato nella Figura 3.3. La distribuzione è multi-modale⁴, infatti si rilevano due picchi distinti: i lavoratori dipendenti, che rappresentano in 33.54% degli intervistati e i pensionati con il 35.63%. Gli appartenenti a queste categorie sono coloro che smaltiscono i rifiuti domestici. Anche le casalinghe raggiungono una percentuale rilevante con il 16.46%.

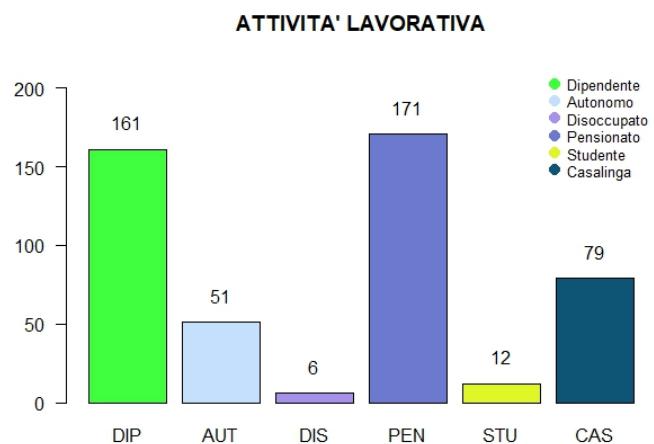


Figura 3.3: Attività lavorativa

Composizione nucleo familiare Ogni intervistato ha indicato la dimensione del proprio nucleo familiare specificando il numero di persone che lo compongono. La variabile è suddivisa in cinque classi (Figura 3.4); la classe con la frequenza più alta è quella con due componenti, quindi nuclei composti da una coppia di persone.

Quartiere L’ultima variabile anagrafica contiene le informazioni sul quartiere di residenza dei cittadini che hanno preso parte al sondaggio. Il quartiere dove risiede la maggior parte degli intervistati è il settimo mentre solo 6 risiedono nel sesto. Questa variabile sarà utile in un secondo momento per stabilire chi di loro risiede nel centro storico del comune.

⁴Distribuzione multi-modale: distribuzione che presenta più di due picchi distinti anche di diversa altezza (Borra and Di Ciaccio, 2008).

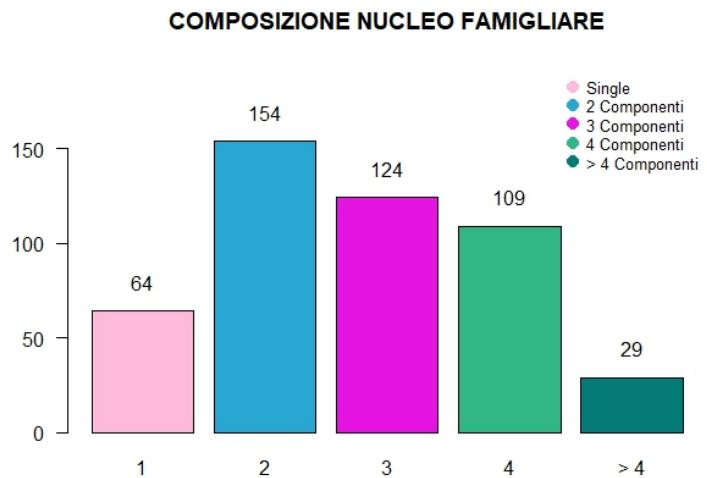


Figura 3.4: Composizione nucleo familiare

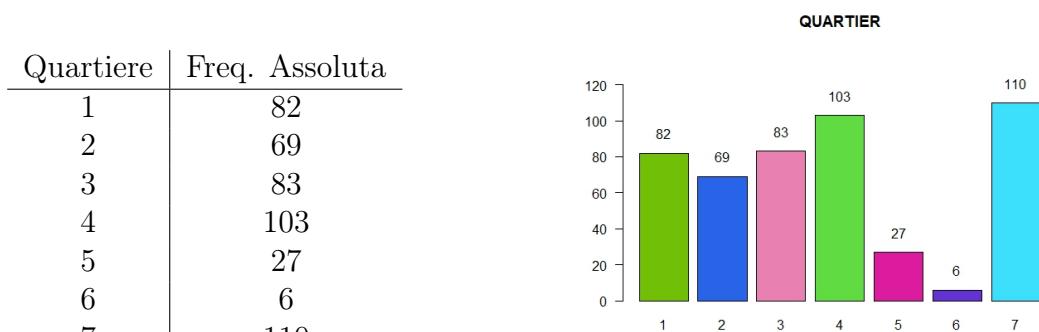


Tabella 3.3: Quartiere

Tabella 3.4: Frequenza distribuzione per quartiere

3.2.3 Informazione sul servizio

Questa sezione del questionario contiene tre domande sul grado di conoscenza che gli intervistati hanno del servizio di raccolta dei rifiuti. Lo scopo è di valutare la quantità e la qualità delle informazioni rilasciate dal comune in merito a questo servizio.

Utilità raccolta differenziata L’obiettivo di questa variabile è capire quale sia, secondo i cittadini, l’utilità del servizio di raccolta dei rifiuti. Ad ogni intervistato sono proposte cinque scelte alternative: (a) Diminuire la quantità di rifiuti da avviare a smaltimento; (b) Ridurre gli sprechi di risorse e materie prime; (c) Far risparmiare il comune; (d) Migliorare il decoro della città; (e) Non ne vedo l’utilità. Le risposte che sono state date il maggior numero di volte sono la prima, 184 volte (38.33%) e la seconda, ben 204 volte cioè il 42.5% (Figura 3.5). In questo caso non c’è alcun dubbio su quale sia il pensiero dei cittadini in merito a questo servizio pubblico: l’obiettivo primario è quindi quello di ridurre le inefficienze nel processo di smaltimento dei rifiuti.

Dubbi sulla raccolta differenziata Un altro elemento d’interesse riguarda gli eventuali dubbi che la popolazione ha in merito alla raccolta dei rifiuti. La Figura 3.6a mostra che più del 90% del campione intervistato dichiara di non aver alcun tipo di dubbio; ciò può significare che le strategie di sensibilizzazione attuate dal comune sono state efficaci.

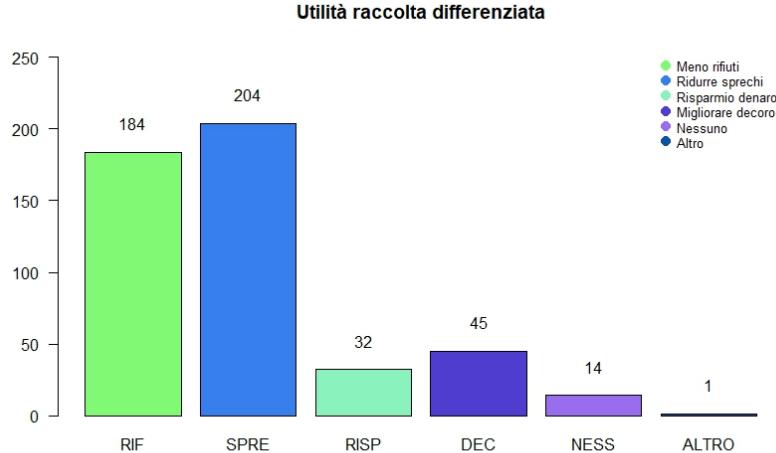


Figura 3.5: Utilità raccolta

Sufficienza delle informazioni Per confermare i dati ottenuti nella precedente domanda, agli intervistati è chiesto un parere personale sulla qualità e quantità delle informazioni divulgate in merito al servizio di raccolta. L’86.46% ritiene che siano sufficienti mentre il 13.54% vorrebbe ricevere maggiori notizie su questo servizio offerto dal comune (Figura 3.6b).

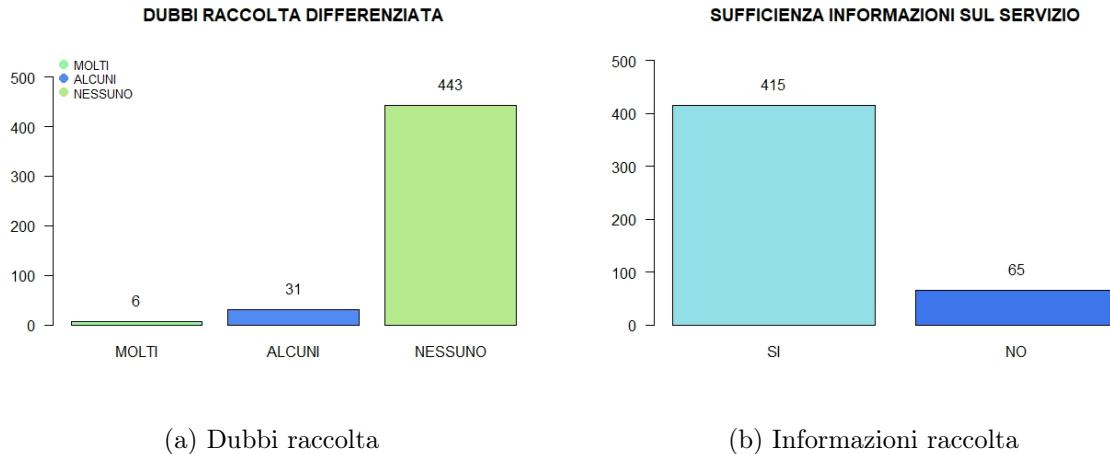


Figura 3.6: Informazioni sul servizio

3.2.4 Ruolo operatori

La terza parte del questionario valuta il lavoro degli operatori quando sollecitati da un suggerimento, reclamo o segnalazione eseguita da un cittadino. Analizzare quest’aspetto del servizio è importante poiché la soddisfazione può risentire di quei casi in cui gli operatori o non riescano a soddisfare una richiesta o, ancor peggio, non la prendano neanche in considerazione.

Fatta eccezione per la variabile sulla professionalità degli operatori, le altre sono escluse dall’analisi perché poste ad un numero troppo esiguo di cittadini, cioè solo a coloro che abbiamo inviato almeno un comunicazione.

Segnalazione, reclami, suggerimenti Sul totale delle 480 osservazioni considerate, solo 47 hanno inviato una segnalazione, un reclamo o un suggerimento. Solo il 9.8% quindi ha svolto un'azione costruttiva per cercare di migliorare il servizio. La quasi totalità dei cittadini, il restante 90.2%, per qualche ragione non ha mai inviato nessuna comunicazione: alcuni potrebbero essere pienamente soddisfatti del servizio, altri potrebbero non conoscere le modalità con cui effettuarle o i destinatari di questo tipo di comunicazioni, ecc.

Solo agli intervistati appartenenti a quel 10% sono state sottoposte delle domande di approfondimento.

Ufficio destinatario Una delle informazioni rilevanti per l'amministrazione comunale è l'ufficio destinatario di segnalazioni, reclami, suggerimenti.

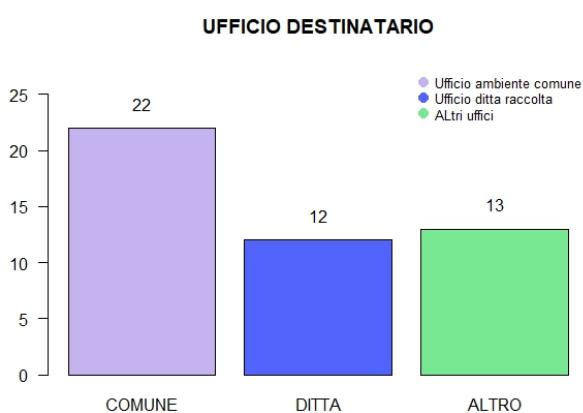


Figura 3.7: Ufficio destinatario

comunicazione ad altri uffici.

L'utilità è verificare se l'ente al quale le persone si rivolgono sia corretto ed eventualmente se ce ne sia uno a cui si rivolgono più frequentemente.

Le risposte sono contenute nella Figura 3.7. Di tutte e 47 le persone che hanno inviato una segnalazione, la maggior parte dei esse, circa il 46.81%, si è rivolta all'ufficio del comune mentre il 25.53% all'ufficio della ditta incaricata della raccolta dei rifiuti. Il restante 27.66% ha invitato la

Risposta ricevuta Non tutti i cittadini hanno ricevuto una risposta da parte dell'ufficio al quale hanno inviato una comunicazione; infatti il 36.17% di loro non ha ricevuto alcun *feedback*. La mancata risposta ad una segnalazione può essere un elemento decisivo per la soddisfazione del cittadino che inviando la segnalazione ha dimostrato di voler contribuire al miglioramento del servizio comunale.

Velocità intervento La velocità con cui gli operatori sono intervenuti per rispondere alle segnalazioni è contenuta nella Figura 3.8a da cui si evince che gli operatori hanno risposto tempestivamente solo 24 volte su 47 mentre per ben 12 volte non si sono presentati.

Risultato intervento Oltre alla tempestività dell'intervento, l'amministrazione comunale chiede ai cittadini di esprimere un giudizio in merito all'efficacia dell'intervento degli operatori. Un dato sicuramente positivo è che ben 27 interventi su 47 sono stati giudicati efficaci mentre solo otto sono giudicati inefficaci (Figura 3.8b).

Professionalità operatori L'ultima domanda è rivolta a tutti gli intervistati, ai quali si chiede di esprimere un giudizio complessivo sul lavoro svolto dagli operatori in una scala di cinque valori: 'Eccellente' e 'Buono' sono considerati positivi, 'Scarso' e 'Insufficiente'

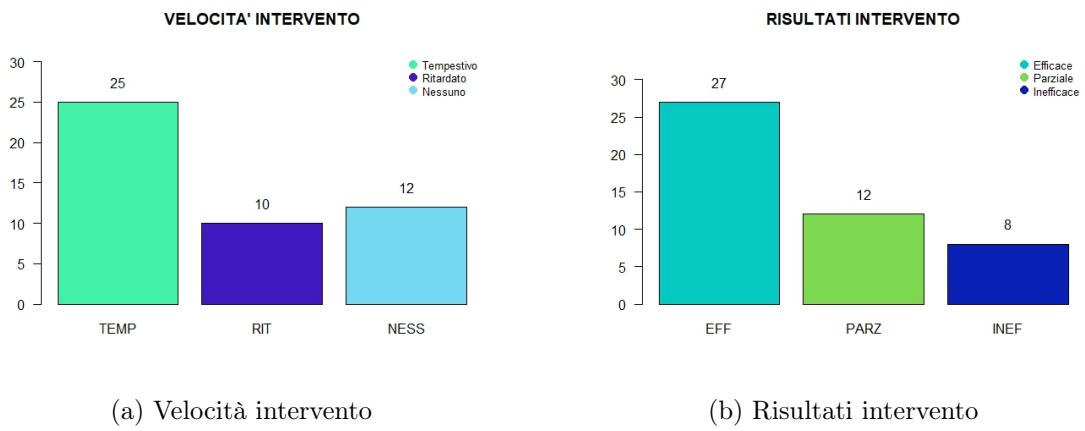


Figura 3.8: Intervento operatori

sono negativi, 'Sufficiente' è un voto neutrale.

Nel complesso il giudizio è decisamente soddisfacente poiché il 75% degli intervistati ha espresso un giudizio positivo, rispettivamente 11.04% per 'Eccellente' e 63.96% per 'Buono', mentre i voti negativi sono stati solo 24, cioè il 5%. I dati relativi alle frequenze assolute sono riassunti nella Figura 3.9.

Le persone che non hanno risposto o che non sono state in grado di valutare gli operatori

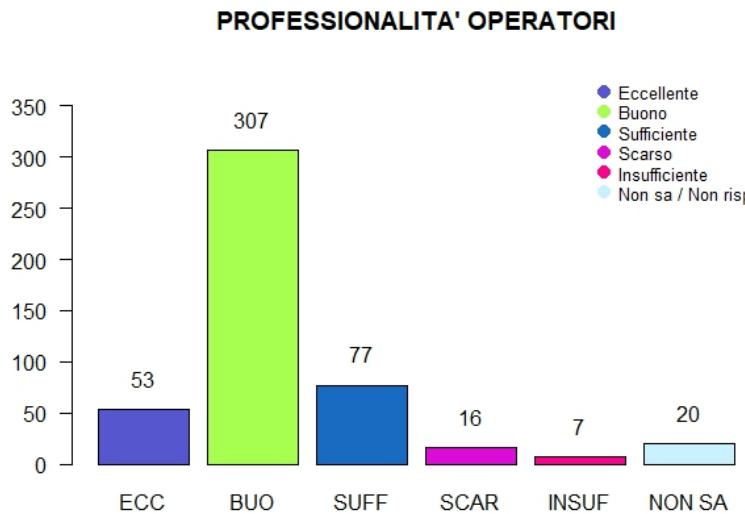


Figura 3.9: Professionalità operatori

sono state 20; probabilmente non hanno mai avuto a che fare con una di queste figure e per tanto non sono capaci di rispondere a questa domanda.

3.2.5 Valutazione del servizio

Questa sezione rappresenta la parte centrale del questionario su cui si concentreranno poi le analisi e le considerazioni finali. Si chiede ad ogni intervistato di esprimere una valutazione sull'importanza e la soddisfazione di alcuni aspetti del servizio di raccolta dei rifiuti urbani. Sono previste due scale di valutazione diverse:

		Importanza		Soddisfazione	
Aspetto	Fattore	Media	Dev.Std.	Media	Dev.Std.
Raccolta Bisettimanale	N. Giorni Sett.	3.918	1.035	7.365	2.096
	Regolarità Servizio	3.837	0.892	8.143	1.690
	Comport. Operator.	3.458	0.848	8.102	1.751
	Resistenza Sacchetti	4.025	1.055	5.394	2.320
Isole Ecologiche	Distribuz. Isole	3.977	0.937	6.794	1.946
	Capienza Camp.	3.850	1.053	6.133	2.055
	Pulizia Isole	3.995	0.984	5.889	2.025
Container	Disponibilità	3.977	0.937	6.794	1.946
	Comodità Orario	3.981	1.035	6.614	2.006
	Praticità Conferim.	3.806	0.972	7.042	1.868
Ecostazioni	Distribuzione	3.856	1.039	7.012	1.858
	Comodità Orario	4.050	1.004	6.502	1.924
	Praticità Conferim.	3.819	0.994	6.931	1.910
Spazzamento e Pulizia Strade	Frequenza	3.798	0.970	6.842	1.956
	Pulizia Strade	4.073	0.926	6.856	2.009
	Distribuz. Cestini	4.006	1.012	6.185	2.001

Tabella 3.5: Valutazione fattori

1. Importanza: composta dai valori compresi tra 5 ed 1, prevede di assegnare 5 al fattore ritenuto più importante e dare un voto da 4 ad 1 ai restanti;
2. Soddisfazione: composta dai valori compresi valori tra 1 e 10, dove il 10 rappresenta la soddisfazione totale, il 6 la sufficienza e così via.

I voti non sono esclusivi quindi lo stesso punteggio può essere assegnato a due fattori diversi. Un gruppo di esperti ha suddiviso il servizio di raccolta in cinque macro categorie, chiamate 'aspetti':

- (a) Raccolta bisettimanale porta a porta di secco ed umido;
- (b) Isole ecologiche: punti adibiti alla raccolta di carta, vetro, lattine e plastica;
- (c) Container: aree adibite alla raccolta dei rifiuti ingombranti/vegetali presso i quartieri;
- (d) Ecostazioni: aree appositamente attrezzate;
- (e) Spazzamento e pulizia delle strade.

A loro volta, ciascuno di questi elementi è stato scomposto ulteriormente in 'fattori' aggiungendo un ulteriore livello di dettaglio all'analisi della soddisfazione dei cittadini.

Valutazione fattori Per ogni fattore sono calcolati i dati di sintesi relativi a media e deviazione standard in modo da poter compiere un giudizio riassuntivo della soddisfazione e dell'importanza. I dati sono riassunti nella Tabella 3.5. Sotto il profilo dell'importanza, tutti i fattori hanno registrato mediamente lo stesso grado di valutazione, che si aggira attorno a 3.9 con una deviazione standard di circa 1. La media più bassa è quella del

'comportamento degli operatori' mentre la più alta è la 'pulizia delle strade'. Per quanto riguarda la soddisfazione invece, i voti sono, in proporzione più bassi visto che le medie si aggirano attorno al valore 6.8 con una deviazione di circa 2 punti. La variabile che ha registrato i punteggi più bassi è la 'resistenza dei sacchetti' mentre quella ritenuta più soddisfacente è la 'regolarità del servizio'.

Queste rappresentano le potenziali variabili risposta da analizzare durante le valutazioni sui possibili scenari, dopo aver appreso la struttura ed i parametri della rete.

Valutazione aspetti Agli intervistati è stato chiesto di dare un voto globale sull'importanza dei cinque aspetti attribuendo '5' all'elemento più importante e '1' a quello meno rilevante. Dalle risposte ottenute si può evincere che la 'raccolta porta a porta' è

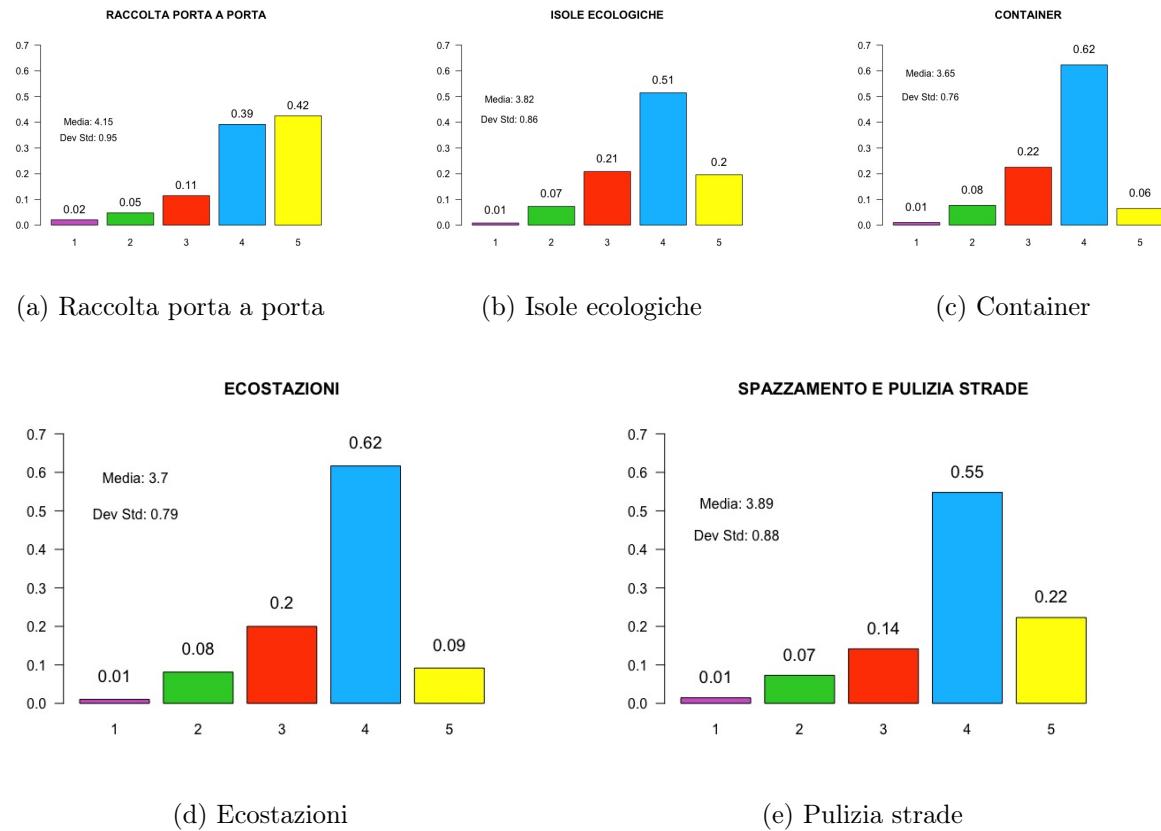


Figura 3.10: Importanza globale degli aspetti

l'aspetto più importante del servizio di pulizia e smaltimento dei rifiuti; esso ha ottenuto il numero di valutazioni positive ('4' e '5') più elevato. Le restanti voci hanno una distribuzione molto simile che ha come classe modale '4'. Anche se ci sono buoni margini di miglioramento, il comune ha registrato, per ogni elemento, un numero davvero esiguo di giudizi insoddisfatti (Figura 3.10).

Per motivi di praticità, d'ora in avanti le variabili potrebbero essere richiamate utilizzando il nome assegnatogli all'interno della banca dati al posto della loro descrizione; in appendice è riportata la tabella 8 contenente la corrispondenza tra il nome e il significato di ogni variabile.

3.2.6 Conoscenza ed utilizzo dei servizi complementari

Il comune mette a disposizione molteplici servizi accessori ed è interessato a raccogliere informazioni sul loro effettivo utilizzo e conoscenza da parte dei proprio cittadini. A ciascun intervistato è chiesto se sia a conoscenza ed eventualmente utilizzi questi servizi. Le risposte sono riassunte nella Tabella 3.6. Osservando le risposte si può concludere che

Servizio	Non Conosco	Conosco	Utilizzo
Composer domestici	114	197	169
Kit micro-raccolta amianto	437	41	2
Raccolta vetro	351	101	28
Raccolta carta	399	69	12
Raccolta toner	364	78	38
Raccolta pile	51	18	411
Raccolta farmaci	14	21	445
Pulizia parchi giochi	247	233	0
Pulizia tombini	147	333	0
Kit bidone dell'umido	23	57	400

Tabella 3.6: Altri servizi di raccolta

tra i servizi elencati sono pochi quelli di cui realmente i cittadini usufruiscono abitualmente, tra cui il 'kit del bidone dell'umido', la 'raccolta pile' e la 'raccolta farmaci'. La maggior parte degli intervistati non è a conoscenza dell'intera gamma di servizi offerti per la raccolta dei rifiuti urbani e solo una piccola parte non li utilizza anche se ne ha notizia. Il comune potrebbe aumentare la soddisfazione dei proprio cittadini in due fasi successive: una prima fase informativa ed una seconda in cui i cittadini possano entrare in contatto con questi servizi.

Anche questo gruppo di variabili sono escluse dalla rete Bayesiana poiché le informazioni che contengono si ritengono non significative nella soddisfazione dei cittadini in questo contesto.

3.2.7 Livello di contribuzione

I fondi per finanziare il servizio di raccolta dei rifiuti derivano dalle tasse versate dai cittadini ogni anno. Risulta fondamentale reperire informazioni su ciò che pensano gli intervistati in merito all'attuale tassazione e le possibili variazioni future.

Conoscenza agevolazione Come prima cosa, il comune vuole conoscere quante persone siano a conoscenza dell'agevolazione del 20% sulla tassa per i rifiuti nel caso in cui il cittadino effettui il compostaggio domestico (smaltimento in proprio dell'umido). Sull'agevolazione non c'è alcun dubbio, il 75% ne è a conoscenza mentre solo il restante 25% degli intervistati ne ha sentito parlare per la prima volta (Figura 3.11). Si ritiene che l'essere a conoscenza dell'agevolazione non sia influente sulla soddisfazione dei cittadini, pertanto si esclude questa variabile dalle analisi successive.

A seguito di variazioni nella legislazione, il comune è obbligato ad aumentare la tassazione per mantenere inalterato il livello attuale del servizio. L'amministrazione vuole conoscere il pensiero degli intervistati in merito alle due alternative per non incorrere in un aumento del prelievo fiscale a carico dei cittadini.

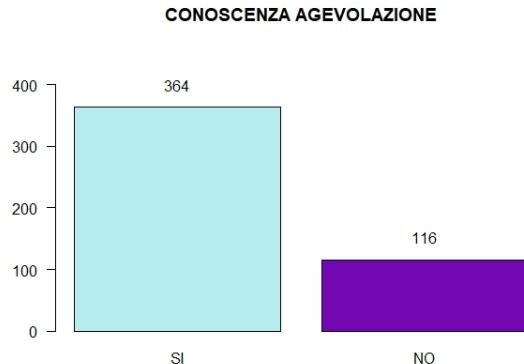
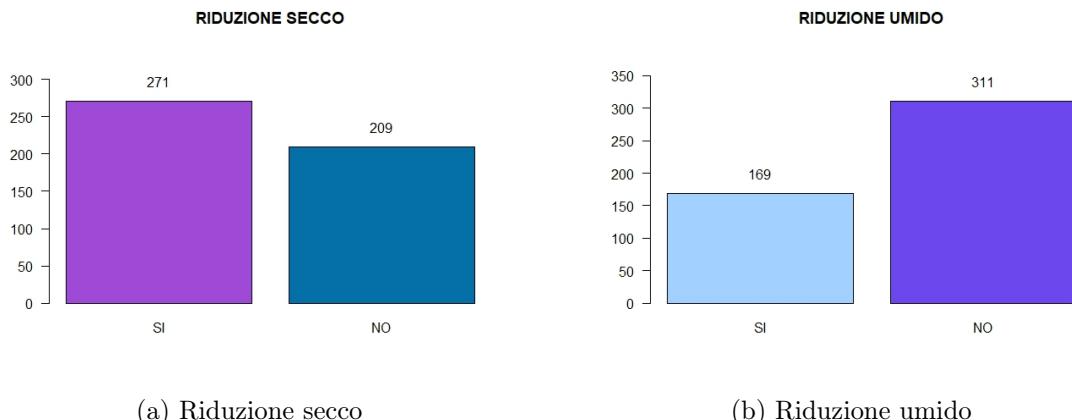


Figura 3.11: Conoscenza agevolazione

Riduzione raccolta secco La prima prevede una riduzione del passaggio della raccolta del secco ad una volta alla settimana. I risultati, 271 voti a favore e 209 contrari, mostrano una leggera approvazione da parte della popolazione verso questa soluzione (Figura 3.12a)

Riduzione raccolta umido La seconda, invece, consiste nell'eliminazione del terzo passaggio di raccolta dell'umido nel solo periodo estivo. In questo caso è evidente come la maggior parte delle persone non sia propensa verso quest'alternativa. Ben 311 persone sarebbero contrarie mentre 169 accetterebbero la variazione (Figura 3.12b).



(a) Riduzione secco

(b) Riduzione umido

Figura 3.12: Propensione alla riduzione della frequenza del servizio

3.2.8 Raccolta con modalità porta a porta nel centro storico

L'ultimo gruppo di domande indaga sulla possibile riduzione di decoro del centro storico dovuta alla raccolta porta a porta di alcune categorie di rifiuti come il secco o l'umido.

Decoro centro storico La prima domanda è rivolta all'intero campione e vuole verificare se la modalità di raccolta porta a porta nel centro storico impatti negativamente sul suo decoro. La frequenza delle risposte date a questo interrogativo sono molto simili tra loro e non si può stabilire quale sia il pensiero dominante nella popolazione. In tutto sono 10 intervistati non hanno saputo dare una risposta a questa domanda; questo può essere dovuto al fatto che non siano residenti nel centro storico o che non lo frequentino.

PERDITA DECORO DEL CENTRO STORICO

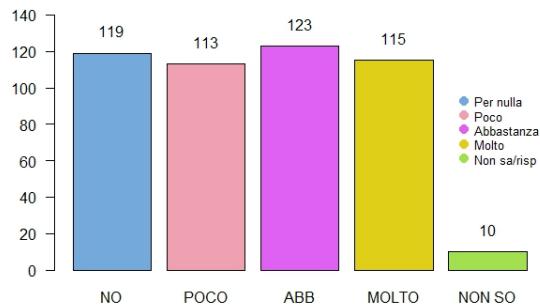
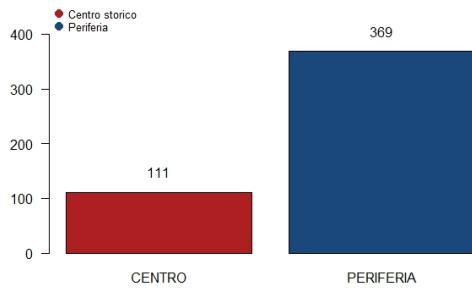


Figura 3.13: Decenza centro storico

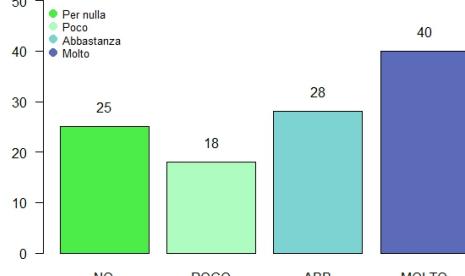
Gli intervistati residenti nel centro storico sono in tutti 111 e rappresentano il 23.13% del campione (Figura 3.14a). Le domande successive riguardano due possibili soluzioni al problema del decoro e sono riservate ai soli residenti nel centro storico che ritengano il servizio un problema per il decoro e che quindi abbiano risposto alla prima domanda (Figura 3.14b) in modo positivo. Gli intervistati in questione sono in tutto 68 e rappresentano il 61.26% dei residenti, quindi la maggior parte degli abitanti del centro storico ritiene la raccolta porta a porta un problema. Essendo state poste poste ad un numero ristretto di intervistati le informazioni raccolte con questi ultimi due quesiti non sono incluse nella rete Bayesiana.

RESIDENTI CENTRO STORICO



(a) Residenti nel centro storico

RESIDENTI INSODDISFATTI



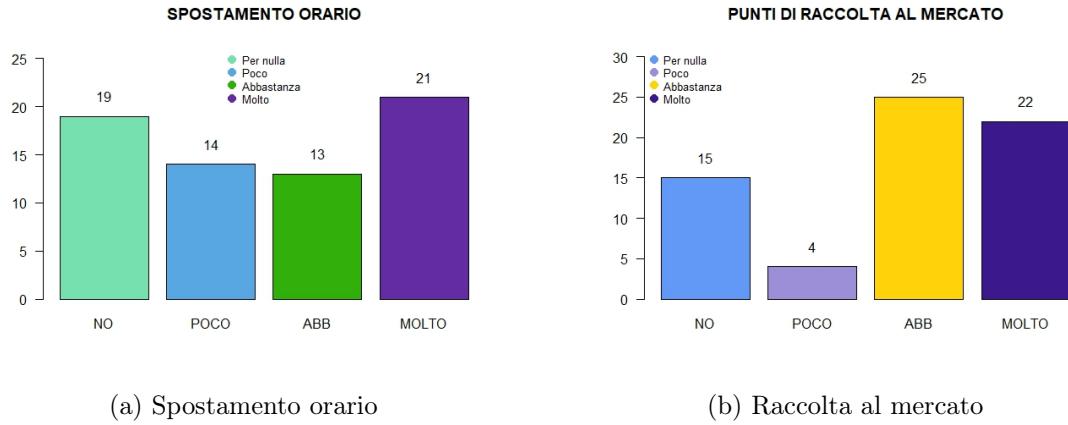
(b) Risposte residenti

Figura 3.14: Informazioni sui residenti nel centro storico

Spostamento orario La prima soluzione prevede lo spostamento dell'orario della raccolta al mattino dalle 8.00 alle 10.00. I risultati sono riassunti nella Figura 3.15a e dimostrano che c'è una parità tra le risposte favorevoli, 13 'abbastanza' e 21 'molto', e le contrarie, 19 'contrarie' e 14 'poco favorevoli'. Per il comune ciò significa che quest'alternativa risulterebbe, secondo i cittadini, ininfluente.

Punti di raccolta al mercato La seconda proposta prevede di conferire i rifiuti anche nei punti di raccolta disponibili nei giorni di mercato. A questa risposta gli intervistati hanno risposto per la maggior parte in maniera favorevole, infatti sono stati registrati 47

voti a favore contro i 19 a sfavore (Figura 3.15b). Attuare questa variazione potrebbe aumentare la soddisfazione dei cittadini.



(a) Spostamento orario

(b) Raccolta al mercato

Figura 3.15: Proposte per preservare il decoro del centro storico

3.2.9 Analisi delle correlazioni

Come precedentemente introdotto, le reti Bayesiane sono ottimi strumenti per rappresentare le relazioni intercorrenti tra le variabili presenti all'interno del sistema analizzato. In statistica i legami più semplici sono descritti dal concetto di correlazione⁵ che misura il grado di linearità tra due variabili. Al contrario, le reti Bayesiane sono i grado di apprendere e rilevare relazioni di qualsiasi tipo. I risultato dell'analisi delle correlazioni saranno utili come termine di paragone con quanto appreso e rappresentato dal modello di BN: ci si aspetta che la rete Bayesiana confermi quanto emerso dal calcolo delle correlazioni e che evidenzi anche i legami più complessi che queste non sono in grado di rilevare.

La correlazione può essere calcolata solo tra variabili di tipo numeriche. Per questo motivo, le uniche variabili che possono essere impiegate in quest'analisi sono quelle relative alla soddisfazione o all'importanza. Per ciascuna coppia di variabili numeriche incluse nella BN è calcolato l'indice di correlazione per misurare la forza della loro relazione.

Definizione 3.1 (Indice di correlazione). *Date due variabili statistiche X ed Y , l'indice è definito come:*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (3.1)$$

dove ρ_{XY} è la covarianza tra X ed Y e σ_X , σ_Y sono rispettivamente le deviazioni standard delle variabili.

Il coefficiente assume valori compresi tra $-1 \leq \rho_{XY} \leq +1$, dove gli estremi dell'intervallo rappresentano la massima correlazione, positiva e negativa, tra le variabili; al contrario valori prossimi allo 0 sono indicatori di un legame debole.

L'analisi è svolta mettendo a confronto i dati di ciascun 'aspetto' con quelli dei 'fattori' in cui è scomposto. Ragionevolmente si può ipotizzare che importanza e soddisfazione di ogni fattore siano correlate, come nel caso della 'raccolta bisettimanale' (Figura 3.16a).

⁵Correlazione: misura della relazione lineare fra due o più variabili. Si dice positiva quando al crescere dell'una crescono anche le altre variabili; viceversa si dice che la correlazione è negativa.

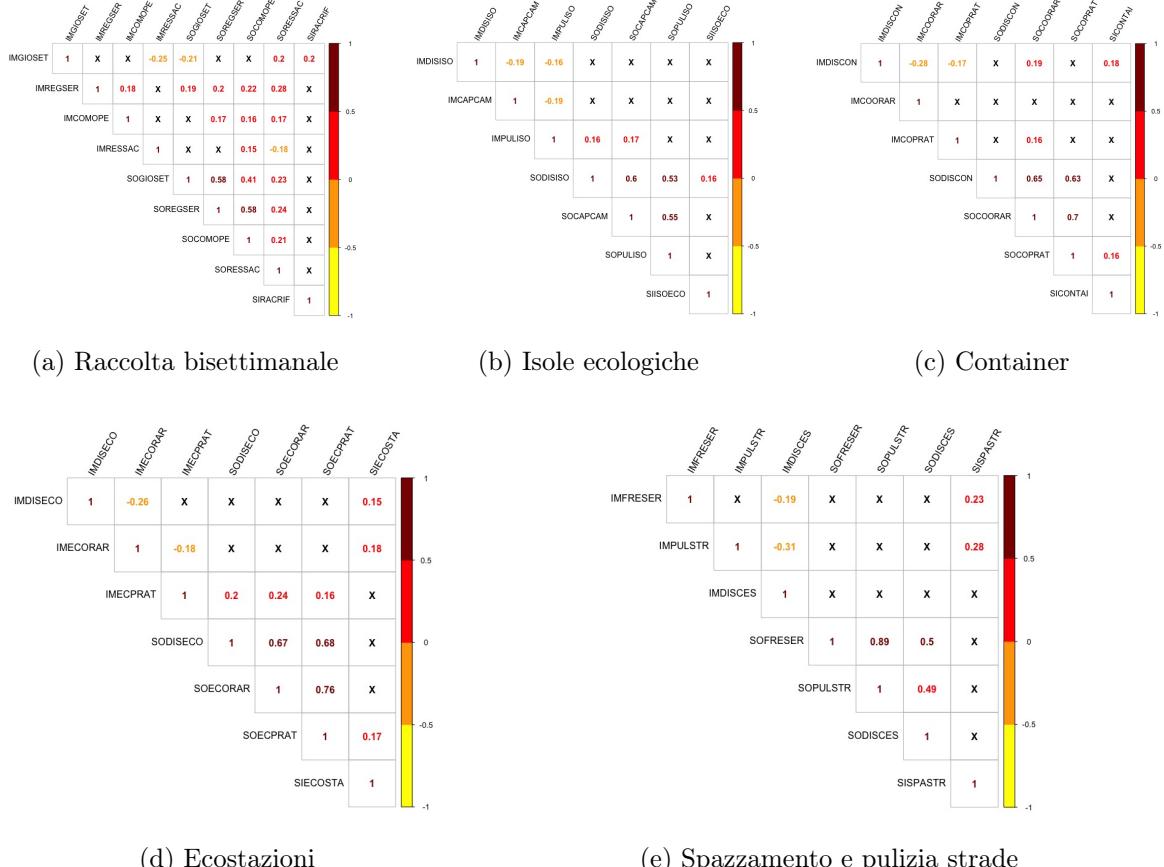


Figura 3.16: Correlazioni

I risultati, invece, mostrano che non sempre esiste una relazione lineare tra questi due elementi come, ad esempio, per lo 'spazzamento e pulizia delle strade' (Figura 3.16e). Un discorso analogo riguarda l'importanza dei cinque aspetti: ognuna dovrebbe essere correlata all'importanza dei singoli fattori in cui è diviso ogni aspetto, ma ciò non è sempre vero. Nel caso delle 'isole ecologiche' (Figura 3.16b), l'importanza globale non è legata all'importanza di nessun fattore mentre quella dello 'spazzamento e pulizia delle strade' è correlata a quella di due fattori su tre.

Il gruppo di variabili sulla soddisfazione è quello che contiene il maggiore numero di relazioni rilevanti, identificate da indici di correlazioni significativi. Questo induce a pensare che nella mente dei cittadini le soddisfazioni formano una rete di interconnessioni molto fitta.

Esistono relazioni anche tra variabili appartenenti ad aspetti diversi. Ne sono un esempio le informazioni sui due orari, quello dei container e quello delle ecostazioni che risultano essere correlati positivamente.

Per concludere, dall'analisi delle correlazioni è emerso che sussistono diversi legami tra le variabili che possono essere spiegati attraverso una funzione lineare; tuttavia mancano le relazioni che probabilmente sono descritte da funzioni più complesse. Ci si attende, quindi, che la rete Bayesiana evidenzi anche le relazioni non spiegate dalle correlazioni.

3.2.10 Alcune considerazioni finali

Durante la presentazione delle variabili, è stato specificato quali possano essere escluse dal modello in quanto non particolarmente informative. Esse sono di seguito riassunte:

- Risposte della sezione "Segnalazioni, reclami, suggerimenti" ad eccezione di quella relativa alla professionalità degli operatori;
- Risposte della sezione "Centro storico" ad eccezione di quella sul decoro del centro storico;
- Risposte della sezione "Altri servizi";
- Le risposte della domanda sulla conoscenza dell'agevolazione.

Dopo aver presentato singolarmente le variabili contenute nella banca dati si ritiene che non tutte abbiano un collegamento, più o meno forte, con la soddisfazione dei cittadini. Il risultato è un *database* privo di NA composto da 480 osservazioni e 50 variabili. Le informazioni in esso contenute sono utilizzate nella fase successiva per apprendere la struttura e i parametri della rete.

3.3 Apprendimento della struttura

Nella prassi è molto frequente che la struttura della rete non sia nota, motivo per cui sono stati sviluppati numerosi algoritmi di apprendimento automatico a partire da un insieme di dati iniziali. Il *database* ottenuto al termine della fase precedente è stato elaborato da diversi algoritmi di *structure learning* appartenenti sia alla famiglia degli approcci *score-based* sia *constraint-based*. I grafi ottenuti sono stati confrontati per individuare il più rappresentativo del contesto analizzato. La rete scelta è quella ottenuta applicando l'algoritmo *Hill climbing* e lo score *Akaike information criterion*⁶ nella sua versione ibrida. Per "ibrida", s'intende che le variabili sono state classificate come categoriali e come numeriche discrete; questa distinzione ha permesso di creare una rete più completa ed esaustiva dal punto di vista delle relazioni individuate.

Gli algoritmi di apprendimento restituiscono la struttura che, secondo le regole con cui sono implementati, meglio descrive le relazioni tra le variabili del sistema. Un arco fra due nodi evidenzia la presenza di un legame statistico tra le due variabili corrispondenti. Si nota che la creazione di un arco si basa sulle relazioni statistiche intrinseche nei dati raccolti in partenza. Questo comporta che una BN potrebbe contenere relazioni di discutibile significato se valutate nel contesto reale che si sta analizzando. Per questo, a seguito dell'apprendimento automatico della rete a partire dai dati, molto spesso segue una valutazione critica del significato intrinseco che essa produce. Chi esegue l'analisi deve immaginare di vestire i panni del destinatario dei risultati (in questo caso il comune) e valutare il risultato dell'apprendimento, magari coadiuvato dai cosiddetti "esperti del settore". Per comprendere meglio il problema si riporta un esempio: l'algoritmo di apprendimento ha rilevato un legame significativo tra la il nodo genitore SORESSAC e il figlio SIISOECO; è ragionevole valutare come poco realistica la condizione per cui l'importanza globale data dai cittadini alle isole ecologiche possa in qualche modo essere

⁶Lo score applicato è sempre Akaike information criterion (AIC), solo che questa versione è specifica per l'elaborazione di variabili continue.

legala alla loro soddisfazione sulla resistenza dei sacchetti della raccolta dell’umido. In questi casi, le tabelle di probabilità condizionata sono uno strumento utile da consultare per valutare le relazioni tra i nodi.

Per correggere eventuali situazioni analoghe, ci sono due strade possibili:

- Eliminare questi archi prima di procedere ad eseguire il processo di inferenza;
- Mantenere inalterata la rete prestando attenzione a tener conto di queste relazioni nelle conclusioni.

La posizione assunta in questa tesi è la seconda. La motivazione nasce dal fatto che questi legami, anche se non attinenti alla realtà, sono il risultato prodotto dall’algoritmo il cui obiettivo è stabilire quale sia la struttura che meglio descriva il *database*. Inoltre, questi archi potrebbero fungere da collegamento indiretto tra nodi che non sono vicini.

La struttura appresa è troppo complessa dal punto di vista pratico per essere elaborata con chiarezza, specialmente in fase d’inferenza probabilistica. Nel tentativo di semplificarla è posto un limite, di volta in volta diverso, al numero massimo di genitori per ogni nodo. La rete prodotta fissando il limite a massimo tre nodi genitori è risultata essere il miglior compromesso tra chiarezza, semplicità e completezza delle relazioni individuate. La struttura definitiva è rappresentata nella Figura 3.17, dove sono evidenziati in rosso i nodi *target* del caso studio analizzato, ovvero le soddisfazioni. Per semplificare l’individuazione dei legami diretti identificati per ogni nodo, nella Tabella 3.7 è riportato l’elenco di tutte le variabili considerate nell’analisi e i rispettivi genitori. Inoltre per evidenziare le diverse macro categorie a cui appartengono, si è proposta una caratterizzazione attraverso diversi colori. Il grafo è composto da 50 nodi, ognuno rappresentante una delle variabili considerate nell’analisi. L’algoritmo *Hill climbing* ha rilevato un totale di 123 relazioni dirette, identificate dagli archi. Una delle prime cosa che saltano all’occhio è l’assenza di nodi isolati, quindi ognuno di essi ha almeno un figlio o un genitore; questo porta a pensare che nella mente degli intervistati le variabili siano interconnesse. In particolare ci sono solamente due nodi radice (DUBBDIFF e PROFOPER) mentre i nodi foglia sono 5 (UTILDIFF, INFORACC, IMPULISO, SODISCES, SIISOECO).

3.3.1 Analisi degli archi individuati

Per valutare l’idoneità del modello nel rappresentare il sistema, bisogna spostare l’attenzione sugli archi. Un modello deve essere in grado di rappresentare le relazioni, dalle più semplici alle più complesse, che intercorrono tra le variabili del sistema. Partendo dal presupposto che un’intelligenza artificiale non conosce nulla del fenomeno reale ad eccezione dei dati che riceve in *input*, si conclude che per l’algoritmo di apprendimento non esistono relazioni scontate o prevedibili. Osservando la costruzione delle variabili, si può intuire che ci sia un collegamento tra le variabili CLETA (classe d’età) e ATTATTU (attività lavorativa): gli studenti appartengono generalmente alla classe d’età più bassa, i pensionati a quella più alta mentre le restanti categorie di lavoratori si collocano nelle restanti fasce d’età.

Generalmente, le relazioni più semplici, come quella appena presentata, sono descritte da archi che collegano variabili appartenenti allo stesso gruppo, per citarne alcuni:

- TITSTUD → ATTATTU: l’attività lavorativa degli intervistati è collegata al titolo di studi che essi possiedono;

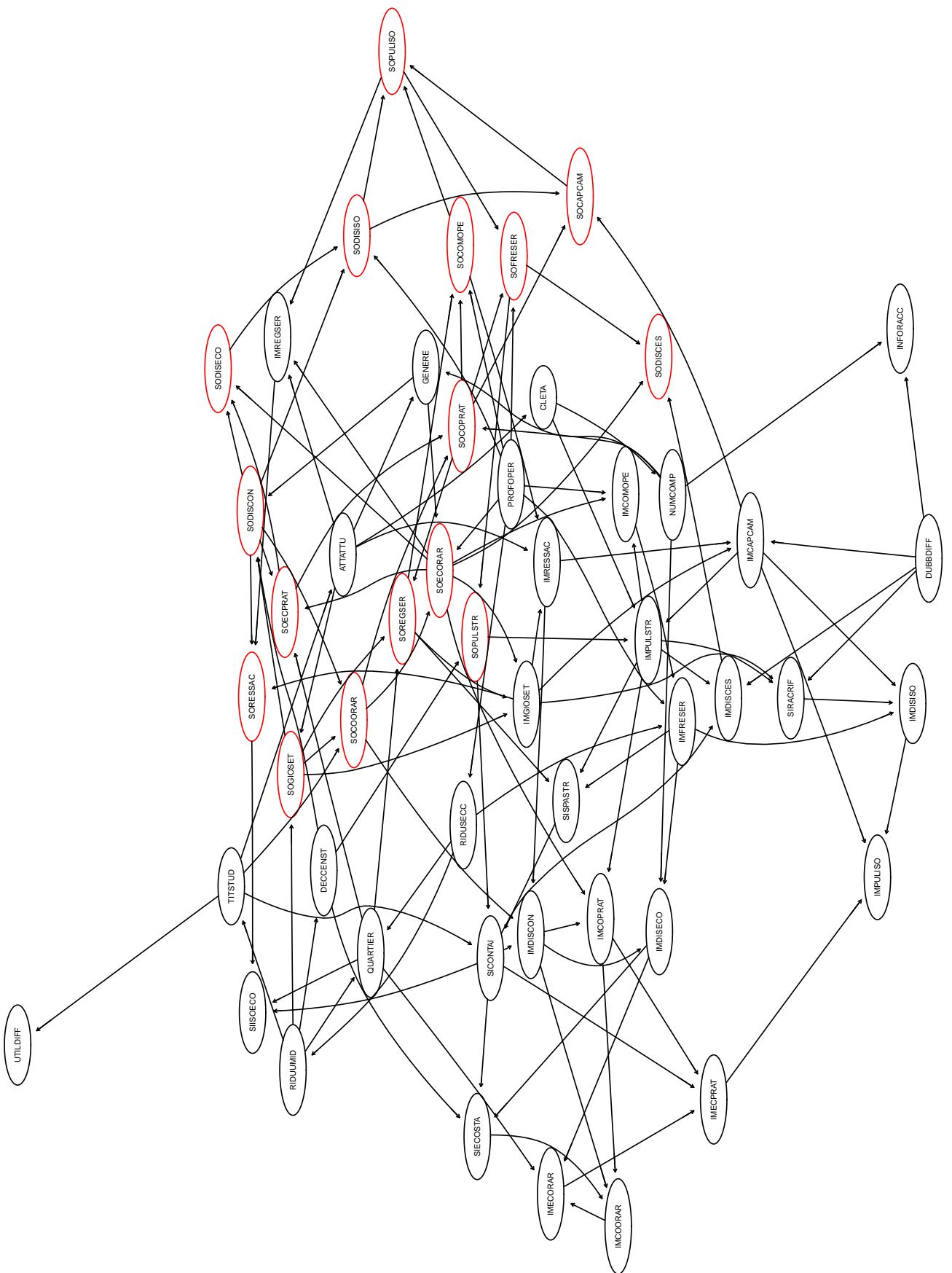


Figura 3.17: Rete Bayesiana

Tabella 3.7: Distribuzione di probabilità condizionata

- NODI -	- GENITORI -	- NODI -	- GENITORI -
UTILDIFF	TITSTUD	SOGIOSET	RIDUUMID
DUBBDIFF	-	SOREGSR	ATTATTU
INFORACC	DUBBDIFF	SOECORAR	
	NUMCOMP	SOPULISO	
PROFOPER	-	PROFOPER	
RIDUSECC	PROFOPER	SOECORAR	
RIDUUMID	RIDUSECC	IMPULSTR	
DECENST	RIDUUMID	IMGIOSET	
SIRACRIF	IMGIOSET	ATTATTU	
	IMPULSTR	SOCOMOPE	
	DUBBDIFF	IMFRESER	
SIISOECO	SICONTAI	IMCAPCAM	
	SORESSAC	SIRACRIF	
	QUARTIER	IMRESSAC	
SICONTAI	TITSTUD	DUBBDIFF	
	SISPASTR	IMGIOSET	
	SOPULSTR	IMECPRAT	
SIECOSTA	SICONTAI	IMCAPCAM	
	DECENST	IMDISISO	
	IMDISECO	SOCOORAR	
SISPASTR	IMPULSTR	SICONTAI	
	IMFRESER	IMRESSAC	
	SOREGSR	IMECPRAT	
TITSTUD	RIDUUMID	IMDISCON	
NUMCOMP	CLETA	SIECOSTA	
ATTATTU	TITSTUD	IMCOPRAT	
GENERE	ATTATTU	IMDISCON	
CLETA	NUMCOMP	IMULSTR	
QUARTIER	ATTATTU	IMDISCO	
	RIDUSECC	IMFRESER	
	RIDUUMID	SOCORAR	
LEGENDA		LEGENDA	
 Soddisfazione fattori		 Soddisfazione fattori	
 Importanza fattori		 Importanza fattori	
 Importanza aspetti		 Importanza aspetti	
 Anagrafica		 Anagrafica	
 Altro		 Altro	

- CLETA → NUMCOMP: il numero di componenti del nucleo familiare dei cittadini è in relazione alla classe d'età a cui appartengono;
- DUBBDIFF → INFORACC: i dubbi che un intervistato ha sul servizio offerto

condizioneranno il suo pensiero riguardo la sufficienza delle informazioni fornite dal comune sulla raccolta dei rifiuti;

- RIDUSECC → RIDUUMID: c'è un collegamento tra la disponibilità a ridurre il numero di passaggi della raccolta del secco e dell'umido;
- SOPULSTR → IMPULSTR: è intuitivo pensare che la soddisfazione e l'importanza di un fattore siano collegate, come nel caso della pulizia delle strade;
- SOGIOSET → SOREGSER: questi nodi appartengono al gruppo delle soddisfazioni e sono legati in modo tale che la soddisfazione sui giorni in cui si effettua, ogni due settimane, il passaggio della raccolta dei rifiuti condiziona quella sulla regolarità con cui gli operatori ecologici rispettano queste giornate.

Gli esempi appena citati, dimostrano la capacità del grafo di rappresentare le relazioni basilari del sistema; se così non fosse, non sarebbero valide le considerazioni successive relative ai legami più complessi.

Le informazioni più interessanti sono nascoste negli archi tra nodi appartenenti a gruppi diversi o che all'apparenza sembrano non essere collegati.

- TITSTUD → UTILDIFF: questa relazione non è del tutto banale e stabilisce che in base al titolo di studi conseguito gli intervistati dichiarano di attribuire al servizio di raccolta dei rifiuti una diversa utilità;
- ATTATTU → SOGIOSET: la soddisfazione sui giorni in cui si effettua ogni due settimane la raccolta porta a porta dei rifiuti è in relazione con l'attività lavorativa svolta dall'individuo; un lavoratore dipendente o autonomo potrebbero essere insoddisfatti dei giorni in cui si effettua la raccolta qualora per motivi di lavoro non potesse essere a casa in quei giorni;
- QUARTIER → IMECORAR: quest'arco stabilisce che l'importanza attribuita all'orario in cui effettuare in conferimento presso le ecostazioni dipende dal quartiere di residenza; esistendo solo tre ecostazioni in tutto il territorio comunale, i residenti più vicino potrebbero attribuire un importanza minore a questo fattore essendo agevolati nel raggiungere il luogo del conferimento;
- PROFOPER → SOFRESER: conoscere il punteggio attribuito agli operatori della raccolta dei rifiuti condiziona la soddisfazione sulla frequenza del servizio di pulizia delle strade;
- SODISECO → SODISISO: questo è un legame tra due nodi che rappresentano la stessa caratteristica, la distribuzione sul territorio comunale, ma di due aspetti diversi, le ecostazioni e le isole ecologiche.

Queste sono solo alcune delle relazioni più complesse individuate dall'algoritmo e sono utili per comprendere la potenzialità delle BN nell'interpretare ed apprendere dai dati informazioni sul fenomeno. Ulteriori legami significativi saranno eventualmente presentati in seguito durante la descrizione dei *Markov blanket* delle variabili *target*.

Confronto con le correlazioni Nel capitolo precedente è stata proposta una rapida analisi delle correlazioni tra le variabili numeriche contenute nel *database*. A differenza delle correlazioni, che esprimo solo il grado di correlazione lineare tra i valori assunti da due variabili, le reti Bayesiane individuano anche legami più complessi di quelli lineari. Ciò che ci si attende dal confronto tra i loro risultati è che la BN da un lato confermi quanto ottenuto dall'elaborazione delle correlazioni e dall'altro ne superi limiti rilevando anche relazioni più complesse di quelle lineari.

Confrontando gli indici di correlazione (Figura 3.16) con gli archi della rete si evince che esattamente il 50% (28 su 56) delle correlazioni, tra l'altro quelle più significative, sono confermate da archi diretti tra i nodi. Nella Figura 3.18 le celle evidenziate in blu indicano le correlazioni confermate dalla rete. Le restanti relazioni lineari, invece, sono inserite

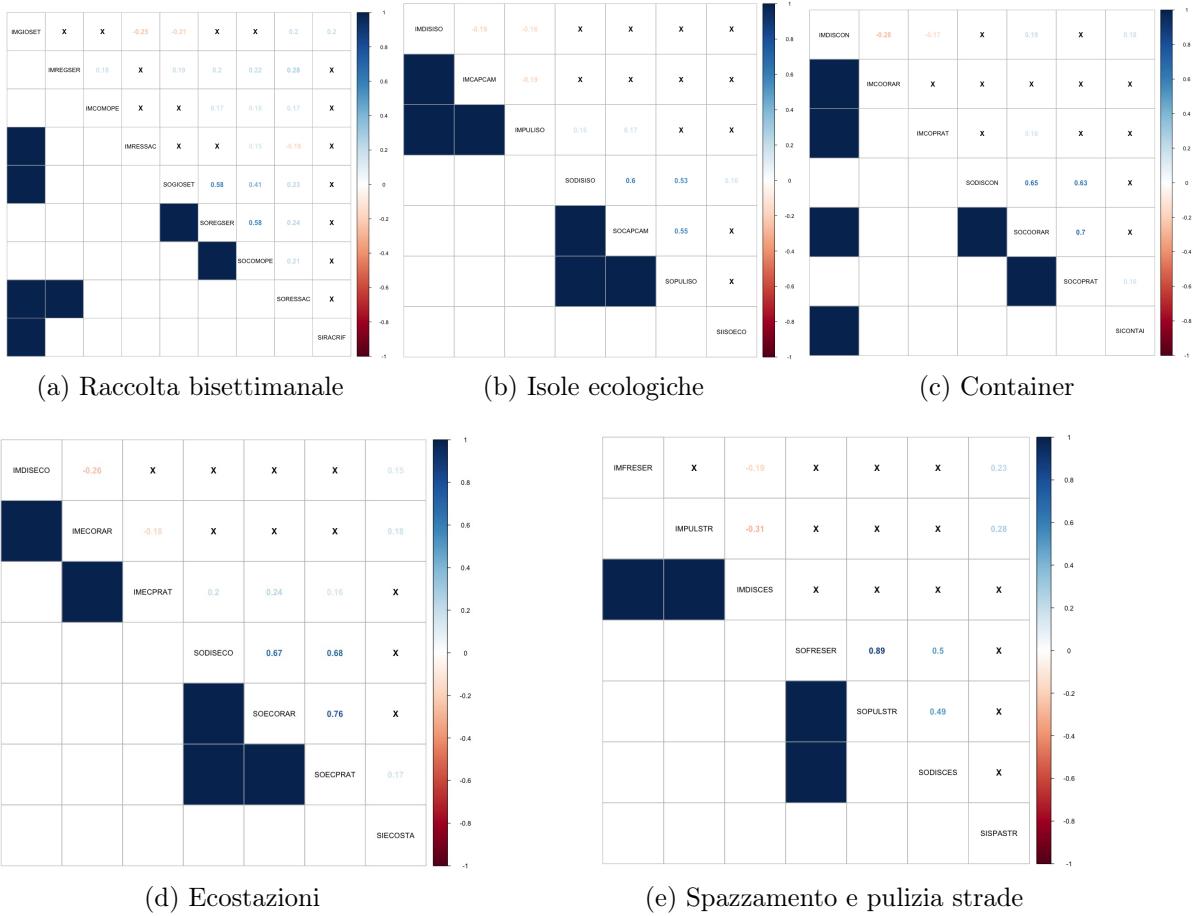


Figura 3.18: Correlazioni confermate dalla BN

nel *Markov blanket* dei nodi quindi esiste una forma di influenza, anche se indiretta, tra le due variabili.

L'algoritmo *Hill climbing* ha superato il limite intrinseco nelle correlazioni introducendo degli archi tra nodi per cui non è calcolato ρ_{XY} . Ad esempio, ci si attendeva che la soddisfazione e l'importanza di uno stesso fattore fossero in qualche maniera legate. Ciò non è emerso dall'analisi delle correlazioni; al contrario, nel grafo relativo alla rete Bayesiana sono evidenti alcuni legami di questo tipo: SOGIOSET → IMGIOSET, IMDISCES → SODISCES, IMCAPCAM → SOCAPCAM. Un discorso analogo riguarda l'aspettativa che ci fosse una relazione tra le soddisfazioni o le importanza appartenenti ad uno stesso aspetto. Anche in questo caso la rete Bayesiana è riuscita ad integrare i risultati

ottenuti con le correlazioni, come: IMCOPRAT→IMCOORAR, SOREGSER→IMGIOSET, ecc.

La conclusione che si può trarre, al termine di questo confronto, è che, come ci si aspettava, la rete Bayesiana non solo conferma quanto appreso attraverso le correlazioni, ma ha compiuto un passo in avanti rilevando relazioni più complesse, rispetto a quelle lineari, portando un maggior livello informativo all'analisi.

3.3.2 Sfera d'influenza delle soddisfazioni

Il fenomeno oggetto di studio, cioè la soddisfazione dei cittadini, è rilevata attraverso 16 variabili: per ogni intervistato nel *database* è memorizzato il grado di soddisfazione per ciascuno dei fattori in cui è suddiviso il servizio di raccolta rifiuti. Non essendoci un solo parametro di riferimento su cui basare l'analisi, sono considerate tutte come variabili "obiettivo"; l'incremento della soddisfazione globale si ottiene massimizzando ciascuna di queste 16 variabili.

Le dimensioni del grafo rendono difficile ragionare con l'intera rete. Il concetto di *Markov blanket* consente di frammentare l'analisi considerando, per ogni nodo, la sotto rete composta dalle variabile da cui non è d-separato. Come introdotto nel capitolo 2, il *Markov blanket* è il sotto insieme di nodi che rende tutto il resto ridondante quando si svolge inferenza su un dato nodo (corollario 2.1). Per ogni nodo *target* è brevemente presentato il sotto grafo su cui hanno influenza nuove informazioni pervenute sul suo valore.

Soddisfazioni giorni settimana - SOGIOSET Il sotto grafo in questione, Figura 3.19a, è composto da 13 nodi collegati da 22 archi direzionali. Al suo interno, le informazioni sul nodo si propagano attraverso:

- 4 variabili anagrafiche: in particolare l'attività lavorativa (ATTATTU) e il quartiere di residenza (QUARTIER);
- 2 variabili relative ad altri fattori: Propensione dei cittadini alla riduzione del passaggio della raccolta dell'umido (RIDUUMID) e la riduzione del decoro del centro storico (DECENST);
- 5 variabili soddisfazione: le più attinenti riguardano la regolarità del servizio di raccolta, appartenente anch'essa all'aspetto della "Raccolta bisettimanale", e agli orari in cui si può conferire nei container (SOCOORAR) e nelle ecostazioni (SOECORAR).

Uno dei punti di forza di questa sotto rete è che le relazioni appena descritte, siano essere dirette o indirette, sono collegabili anche da un punto di vista logico alla soddisfazione dei giorni in cui si effettua la raccolta.

Soddisfazione regolarità del servizio - SOREGSER Questa rete (Figura 3.19b), composta da 11 nodi e 14 archi, descrive la propagazione delle informazioni relative alla soddisfazione sulla regolarità del servizio di raccolta bisettimanale.

Osservando la Figura 3.19b si nota che IMREGSER, cioè il nodo corrispondente all'importanza del proprio fattore, non è tra i vicini del nodo ma è contenuto comunque all'interno della sua area d'influenza, perciò esiste comunque una relazione di loro.

L'unica variabile anagrafica presente è il quartiere (QUARTIER), quindi cittadini residenti in zone diverse hanno espresso un diverso grado di soddisfazione. Senza osservare il

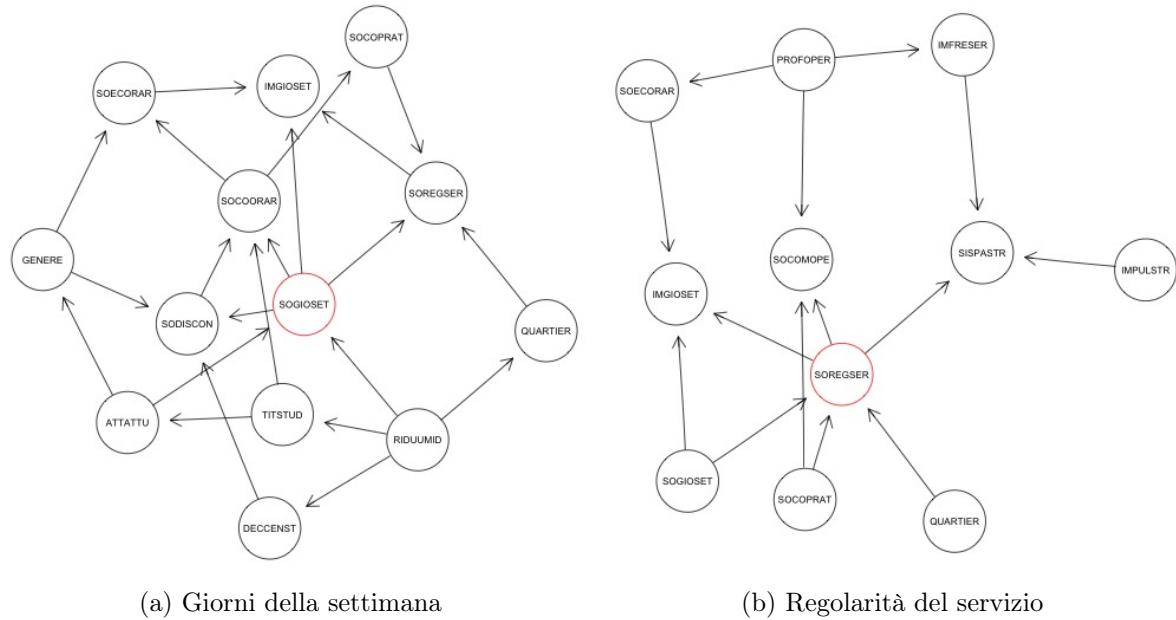


Figura 3.19: Markov blanket

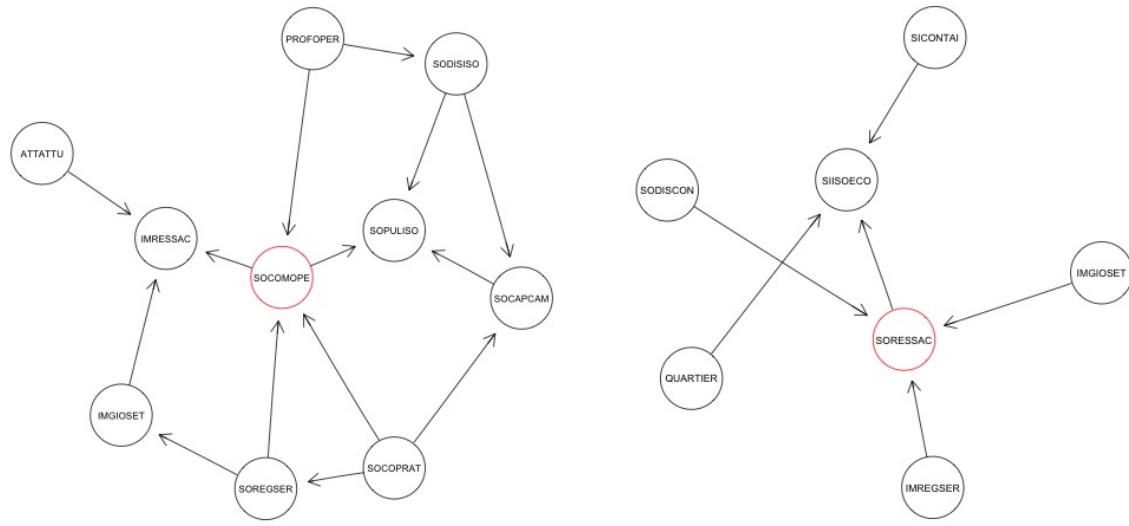
grafo, si può presupporre che la regolarità del servizio sia legata all'attività degli operatori; infatti, in questo grafo sono presenti i nodi relativi alla professionalità degli operatori (PROFOPER) e alla soddisfazione sul loro comportamento (SOCOMOPE), quest'ultimo appartiene anche allo stesso gruppo di fattori del nodo in esame. Con un ragionamento del tutto analogo a quello appena compiuto, non ci si stupisce che siano presenti anche i nodi sulla soddisfazione e l'importanza dei giorni della settimana in cui si effettua la raccolta (SOGIOSET e IMGIOSET): cittadini insoddisfatti dei giorni in cui si compie la raccolta sono insoddisfatti anche della regolarità del servizio.

Soddisfazione comportamento operatori - SOCOMOPE Il *Markov blanket* della soddisfazione sul comportamento degli operatori (Figura 3.20a) comprende 10 nodi e 14 archi, i più rilevanti sono:

- PROFOPER, la professionalità degli operatori che rientra tra i nodi genitore;
- SOREGSER, la soddisfazione sulla regolarità del servizio che appartiene al medesimo aspetto;
- SOPULISO, la soddisfazione sulla pulizia delle isole ecologiche che la quale rientra nei compiti degli operatori ecologici;
- SOCOPRAT, la soddisfazione sulla praticità del conferimento nei container poiché i cittadini sono aiutati dagli operatori in quest'operazione.

L'unica variabile anagrafica che entra in gioco in questo sotto grafo è l'attività lavorativa (ATTATTU).

Soddisfazione resistenza sacchetti dell'umido - SORESSAC La Figura 3.20b rappresenta la rete di dipendenze della soddisfazione sulla resistenza dei sacchetti dell'umido che, tra i quattro fattori in cui è suddiviso l'aspetto della "raccolta bisettimanale", è



(a) Comportamento operatori

(b) Resistenza sacchetti umido

Figura 3.20: Markov blanket

probabilmente quello che nella realtà è più difficile da collegare ad altri elementi. QUARTIER è l'unica informazione sui cittadini contenuta nella sotto rete. Gli unici legami significativi sono quelli con IMGIOSET e IMREGSER, visto che si riferiscono all'importanza rispettivamente dei giorni e della regolarità del servizio con cui gli operatori eseguono la raccolta dell'umido porta a porta. La validità pratica degli archi restanti sarà valutata durante il processo di inferenza, in quanto è difficile, con queste informazioni, attribuirgli un senso logico.

Soddisfazione distribuzione isole ecologiche - SODISISO Osservando la Figura 3.21, la prima cosa che si nota è l'assenza di variabili anagrafiche. Il *Markov blanket* è formato da 9 nodi e i 12 archi direzionali rappresentano legami significativi sia statisticamente sia dal punto di vista pratico. Il collegamento con l'importanza e la soddisfazione della capienza dei cassonetti induce a pensare che al diminuire della direzione, per mantenere inalterata la soddisfazione, si dovrà aumentare la distribuzione delle isole sul territorio. Queste considerazioni saranno verificate durante il processo inferenziale. Anche il nodo SOPULISO (soddisfazione pulizia isole ecologiche) è presente in questa sotto rete, si può concludere che le in-

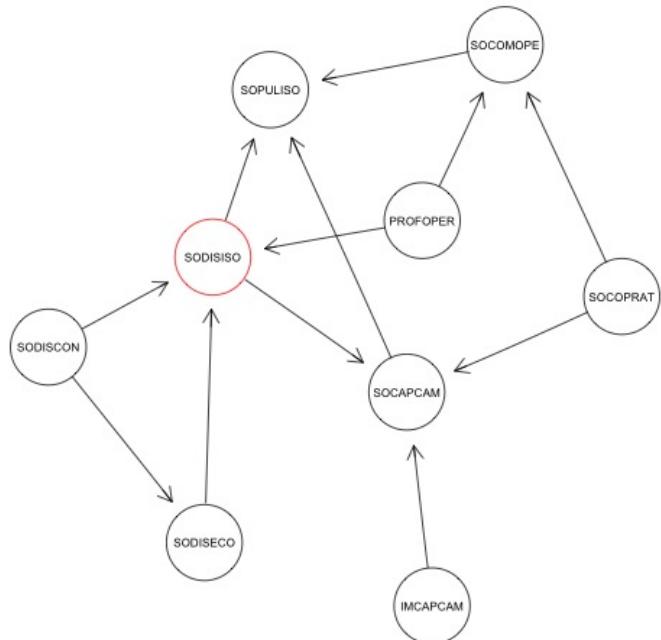


Figura 3.21: Markov blanket - Distribuzione isole ecologiche

formazioni sulla soddisfazione di un qualsiasi fattore delle isole ecologiche influiscano su SODISISO. Il grafo evidenzia un'interessante relazione tra le tre variabili relativa alla distribuzione sul territorio di tre elementi del servizio: le isole ecologiche (SODISISO), le ecostazioni (SODISECO) e i container (SODISCON).

Soddisfazione capienza cassonetti/campane - SOCAPCAM La rete, di ridotte dimensioni rispetto alle precedenti, contiene solo 6 nodi e 7 archi. La corrispondente importanza IMCAPCAM è genitore del nodo, il che presuppone una relazione diretta tra i due. Le relazioni più significative sono quelle che collegano le soddisfazioni dei fattori delle "isole ecologiche": SOCAPCAM, SOPULISO, SODISISO (Figura 3.22a).

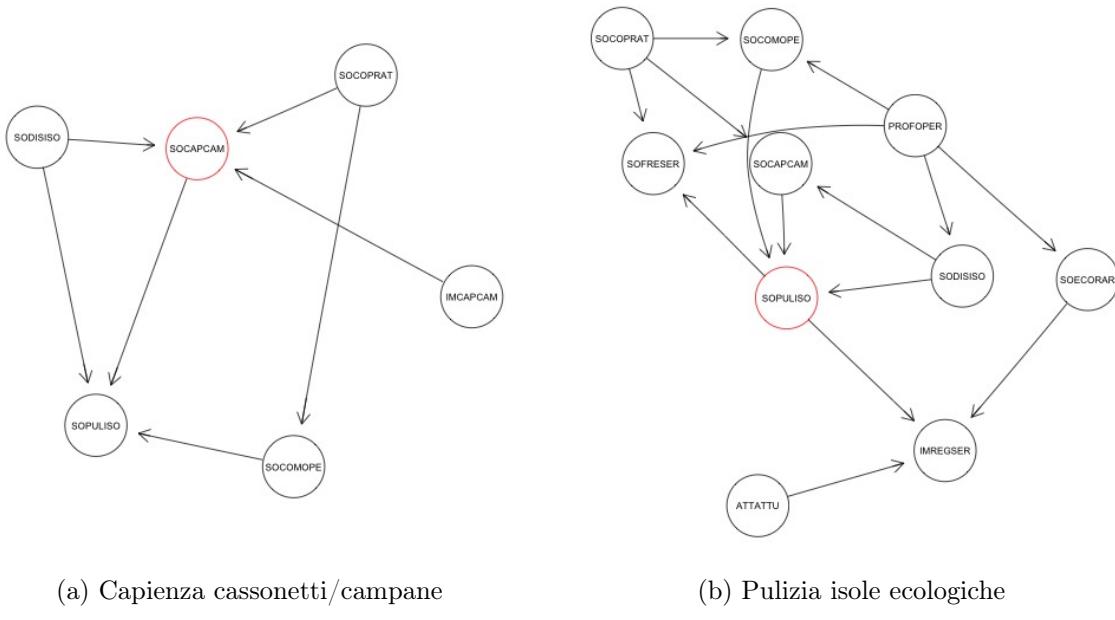


Figura 3.22: Markov blanket

Soddisfazione pulizia isole ecologiche - SOPULISO La seguente considerazione potrebbe risultare ridondante, ma anche nella Figura 3.22b si nota il collegamento tra le soddisfazioni di tutti i fattori delle isole ecologiche. L'attività lavorativa (ATTATTU) è l'unica variabile anagrafica compresa nel *Markov blanket* in analisi. Non sorprende vedere tra i nodi anche due elementi relativi agli operatori, in particolare si tratta dalla valutazione della loro professionalità (PROFOPER) e la soddisfazione sulla loro comportamento (SOCOMOPE). Quest'ultima relazione, per quanto possa apparire ovvia visto che la pulizia del luogo di lavoro rientra tra le normali mansioni di ogni operatore, si ricorda che nulla è scontato per l'intelligenza artificiale che ha appreso la struttura della rete.

Soddisfazione distribuzione container - SODISCON Questa soddisfazione è tra le più influenti: il sotto grafo (Figura 3.23) ottenuto isolando i nodi in cui si propagano le informazioni sulla distribuzione dei container è formato 15 nodi e 24 archi. La distribuzione territoriale dei container chiama in causa diverse variabili anagrafiche: genere (GENERE), titolo di studi (TITSTUD) e quartiere di residenza (QUARTIER). Secondo i dati

in possesso, esiste un'interessante relazione tra il decoro del centro storico (DECENST) e la distribuzione dei container. Alla disposizione dei container sono connesse anche la disposizione delle isole ecologiche e delle ecostazioni, legame già osservato in precedenza.

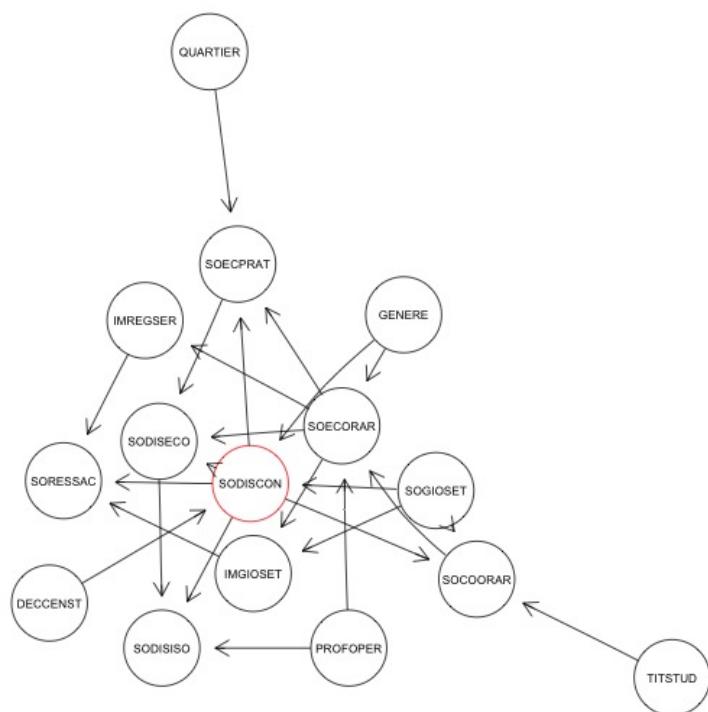


Figura 3.23: Markov blanket - Distribuzione container

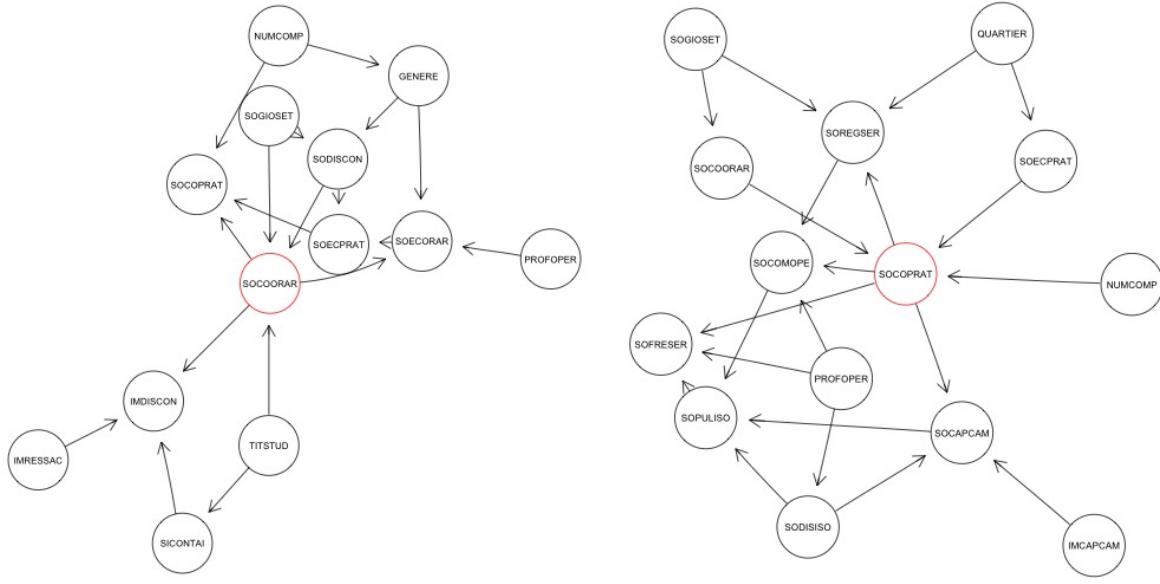
componenti (NUMCOMP). Per qualche ragione gli intervistati con un nucleo familiare diverso non hanno fornito la stessa risposta.

Questo nodo è il primo ad essere collegato all'importanza globale del proprio aspetto, cioè i container (SICONTAI). La soddisfazione sull'orario di conferimento è legata ad altre variabili temporali, quali: la soddisfazione sull'orario di conferimento nelle ecostazioni (SOECORAR) e l'importanza e la soddisfazione dei giorni in cui è svolta la raccolta porta a porta (IMGIOSET e SOGIOSET). All'interno della sotto rete, Figura 3.24a sono presenti anche le soddisfazioni degli altri due fattori in cui sono divisi i container che sono rispettivamente la distribuzione sul territorio comunale (SODISCON) e la praticità del conferimento (SOCOPRAT).

Soddisfazione praticità conferimento nei container - SOCOPRAT Alcune delle relazioni nel grafo rappresentato nella Figura 3.24b, sono già state analizzate e per evitare ridondanze sono solo accennate. Sono presenti le variabili relative agli operatori ecologici che ne descrivono rispettivamente la professionalità globale (PROFOPER) e la soddisfazione sull'operato quando eseguono la raccolta porta a porta (SOCOMOPE). La presenza del nodo relativo all'importanza e alla soddisfazione della capienza dei cassonetti è interessante (IMCAPCAM e SOCAPCAM): nelle isole ecologiche, la capienza delle campane è indicatore della praticità del conferimento quindi due intervistati che attribuiscono un'importanza diversa a questo fattore non hanno lo stesso grado di soddisfazione. Anche in questo caso è confermato il legame con la soddisfazione sulla distribuzione (SODISCON) e agli orari dei container (SOOCORAR). Il quartiere (QUARTIER) e il

Degli altri due fattori, in cui è scomposto l'elemento "container", la distribuzione non è legata alla praticità (SOCOPRAT) ma solo all'orario (SOOCORAR). La spiegazione è che minore è la diffusione dei container e più difficile è per i cittadini arrivarvi, questo comporta che chi non ha molto tempo da dedicare alla raccolta dei rifiuti potrebbe non essere soddisfatto dell'orario in cui sono accessibili i container.

Soddisfazione orario conferimento nei container - SOCOORAR In fase di apprendimento sono emerse relazioni tra questa soddisfazione e tre variabili anagrafiche: titolo di studi (TITSTUD), genere (GENERE) e, sorprendentemente, numero di



(a) Orario conferimento container

(b) Praticità conferimento container

Figura 3.24: Markov blanket

numero di componenti (NUMCOMP), sono i dati anagrafici che intervengono nel processo di propagazione delle informazioni su questo nodo.

Soddisfazione orario conferimento nelle ecostazioni - SOECORAR Il nodo SOECORAR ha un *Markov blanket* composta da ben 21 nodi e 36 archi. Volendo contare il numero dei vicini, dalla Figura 3.25 si contano 7 nodi figli e 3 nodi genitori. Le variabili anagrafiche contenute in questa rete sono 3: genere (GENERE), attività lavorativa (ATTATTU) e quartiere di residenza (QUARTIER); solo il sesso degli intervistati è un nodo genitore di SOECORAR.

Le soddisfazioni riguardanti gli altri due fattori delle ecostazioni cioè la distribuzione (SODISESCO) e la praticità (SOECPRAT) fanno parte dell'insieme dei figli del nodo in esame. Tra i genitori, invece, è presente la soddisfazione sull'orario di conferimento nei container (SOCOOAR), un interessante collegamento l'orario di un altro elemento del servizio. I restanti archi descrivono relazioni interessanti da analizzare in fase di inferenza. Si tratta dei legami individuati con le variabili legate agli operatori e ad alcuni aspetti della raccolta bisettimanale: degli operatori si considerano la valutazione sulla professionalità (PRO-

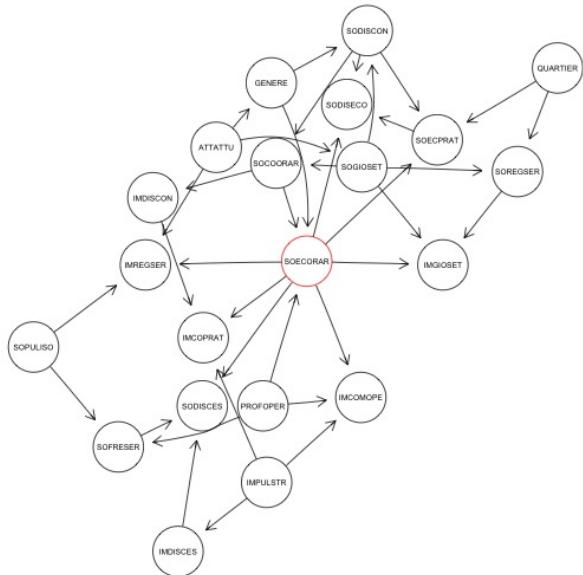


Figura 3.25: Markov blanket - Orario conferimento ecostazioni

FOPER) e la soddisfazione sul loro comportamento durante la raccolta (SOCOMOPE) mentre della raccolta bisettimanale sono presenti l'importanza e la soddisfazione sia dei giorni della settimana (IMGIOSET e SOGIOSET) sia della regolarità del servizio (IMREGSER e SOREGSER).

Soddisfazione distribuzione ecostazioni - SODISECO L'influenza di questo nodo è limitata a 6 nodi connessi da 9 archi, così come descritto nella Figura 3.26a. La distribuzione delle ecostazioni è connessa da un lato alla distribuzione delle isole ecologiche (SODISISO) e dei container (SODISCON), dall'altro all'orario (SOECORAR) e alla praticità (SOECPRAT) del conferimento che sono gli altri due fattori caratterizzanti le ecostazioni. Sebbene le reti sia di dimensioni ridotte rispetto a quelle analizzate in precedenza, descrive le relazioni più essenziali e significative che potessero coinvolgere il nodo SODISECO.

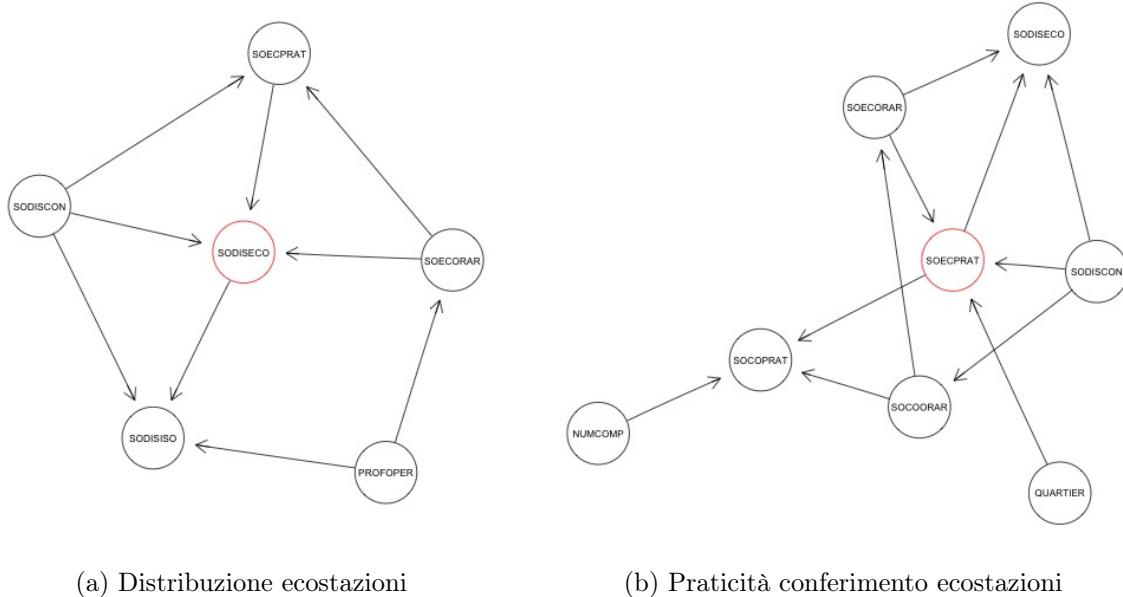


Figura 3.26: Markov blanket

Soddisfazione praticità di conferimento nelle ecostazioni - SOCOPRAT Il *Markov blanket* di questa soddisfazione è formato da 8 variabili connesse da 11 archi. A conferma dell'analisi svolta per le ultime due variabili, anche in questa sotto rete (Figura 3.26b) sono presenti legami con le soddisfazioni sull'orario del conferimento (SOECORAR) e la distribuzione delle ecostazioni (SODISECO). L'influenza si propaga anche attraverso due variabili anagrafiche, il quartiere (QUARTIER) e il numero di componenti del nucleo familiare (NUMCOMP). I nodi restanti rappresentano i tre fattori che caratterizzano i "container", questo presuppone che ci sia un legame tra questi due elementi del servizio.

Soddisfazione frequenza del servizio pulizia strade - SOFRESER La rete in esame contiene lo stesso numero di nodi e di archi, cioè 9 (Figura 3.27). In particolare, tra i figli di SOFRESER ci sono le soddisfazioni dei restanti fattori della "pulizia e spazzamen-

to delle strade": la pulizia delle strade (SOPULSTR) e la distribuzione dei cestini (SODISCES).

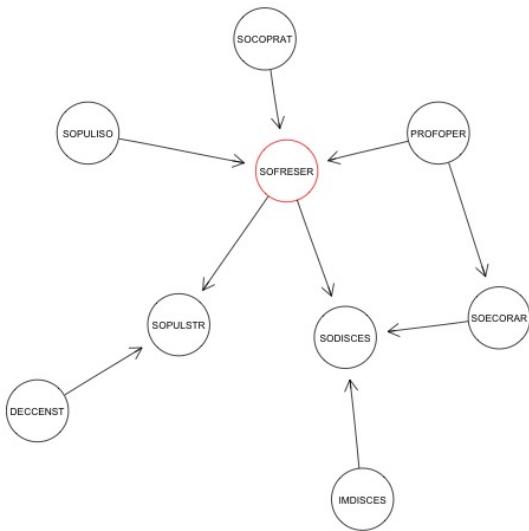


Figura 3.27: Markov blanket - Frequenza servizio pulizia strade

sulla frequenza del servizio (SOFRESER). La presenza di un arco che lega la SOPULSTR alla propria importanza (IMPULSTR) è sicuramente un segnale positivo sull'ideoneità della struttura nel rappresentare il fenomeno. Le dimensione di questa sotto rete (Figura 3.28a) sono gli stessi della precedente, 9 archi e 9 nodi. Uno dei genitori del nodo in esame è il decoro del centro storico (DECENST), verosimilmente, minore è la soddisfazione sulla pulizia delle strade e maggiore sarà l'impressione che il decoro del centro storico diminuisca. Le variabili anagrafiche che influiscono su questa rete sono: titolo di studi (TITSTUD) e la classe d'età (CLETA). Si noti che tra i nodi della rete è compresa anche l'importanza globale attribuita dall'intervistato all'aspetto dello "spazzamento e pulizia delle strade" (SISPASTR).

Soddisfazione distribuzione cestini - SODISCES L'ultimo *Markov blanket* da analizzare è composto da 4 nodi e 3 archi (Figura 3.28b). La SODISCES ha come genitori due variabili che appartengono allo stesso gruppo e sono: la corrispondente importanza del proprio fattore (IMDISCES) e la soddisfazione sulla frequenza del servizio (SOFRESER).

3.4 Apprendimento parametri

Il passo successivo alla definizione della struttura della rete è l'apprendimento di un secondo elemento indispensabile che rappresenta l'aspetto quantitativo del modello, definito dalle tabelle di probabilità condizionata.

Nelle elaborazioni eseguite all'interno di questo processo, si preferisce, generalmente, ricorre all'impiego di metodi Bayesiani piuttosto che alla *maximum likelihood estimation* per due motivi (Nagarajan et al., 2013, pp.104): (a) Le stime Bayesiane sono più semplici rispetto all'MLE, motivo per cui l'inferenza risulta semplice e robusta; (b) Per campioni di piccole dimensioni, i parametri Bayesiani sono più vicini a quelli "veri" e garantiscano

Questi nodi consentono alle informazioni di propagarsi anche all'interno di altri due nodi, rispettivamente: decoro del centro storico (DECENST) e importanza della distribuzione dei cestini (IMDISCES). I dati raccolti attraverso le interviste consentono di rilevare un legame tra la frequenza del servizio e la professionalità degli operatori che svolgono questa mansione (PROFOPER).

Soddisfazione pulizia delle strade - SOPULSTR Delle relazioni con gli altri fattori della pulizia delle strade è confermata solo quella con la soddisfazione

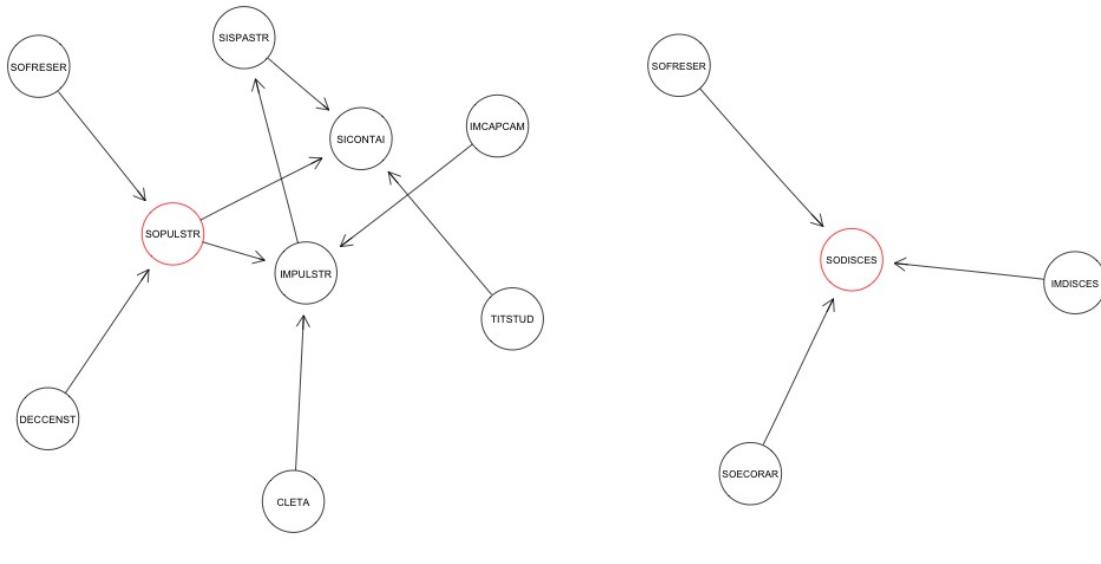


Figura 3.28: Analisi Markov blanket

che nelle tabelle di probabilità condizionata non ci siano dati mancanti.

Nota tecnica: per favorire l'individuazione delle relazioni tra le variabili del sistema, la topologia del grafo è stata appresa servendosi di uno *score* ibrido ("aic-cg"); ciò pregiudica la possibilità di ricorrere ad un approccio Bayesiano in quanto i *software* gratuiti a disposizione per l'apprendimento e l'inferenza delle BN sono limitati. In questo caso, poiché nel sondaggio le variabili sono predisposte per essere raccolte sotto forma di dati categoriali o numerici discreti, è possibile apprendere i parametri della rete considerando come categoriali le variabili che durante l'apprendimento della struttura, mediante l'approccio ibrido, sono state considerate come tali. Attraverso quest'ipotesi, è possibile applicare il metodo Bayesiano ed apprendere le tabelle di probabilità condizionata associate a ciascun nodo della rete.

Elencare ciascuna TPC non è utile ai fini di questa tesi e potrebbe risultare superfluo; nel seguito saranno presentate di volta in volta solo le tabelle più significative.

3.5 Processo d'inferenza

Il risultato delle operazioni compiute fin'ora erano è BN che sotto il profilo sia quantitativo sia qualitativo descrive la soddisfazioni dei cittadini.

Verificata l'idoneità della rete nel rappresentare il fenomeno, si procede con lo svolgimento del processo d'inferenza. Questa fase consiste nell'inserire nel modello delle informazioni riguardanti il valore delle variabili, o evidenze, e monitorare i conseguenti cambiamenti attuati della rete per adattarsi ai nuovi dati introdotti. Attraverso un'analisi mirata dei nuovi parametri si possono trarre importanti conclusioni da utilizzare come supporto all'interno del processo decisionale dell'azienda.

Il software utilizzato per elaborare l'inferenza esatta utilizza l'algoritmo *junction tree*. Il processo è suddiviso in due parti:

1. Inferenza sulle variabili *target*, per determinare i *driver* su cui agire per massimizzare ciascuna soddisfazione;
2. Analisi di scenario, attraverso la manipolazione di più variabili contemporaneamente si possono simulare gli effetti delle strategie aziendali.

Si richiede una certa cautela nell'interpretare i risultati perché nel *database* sono contenute delle variabili, come quelle contenenti informazioni anagrafiche, che il comune non è in grado di controllare né direttamente né indirettamente. Questo significa che non è possibile applicare qualsiasi strategia dovesse emergere dal processo d'inferenza, che richieda di modificare questo tipo di variabili. L'utilità di questi dati risiede nella capacità di interpretare e descrivere il fenomeno più approfonditamente; ad esempio: introducendo l'evidenza $E \rightarrow SOGIOSET = 10$ nella rete, la variazione della CPT di ATTATTU permette di capire quali sono probabilmente le categorie di soggetti maggiormente soddisfatti del campione intervistato.

3.5.1 Identificazione dei driver di soddisfazione

L'obiettivo di questa prima parte dedicata all'inferenza è determinare quali siano i fattori impattano sulla massimizzazione di ciascuna variabile legate alla soddisfazione complessiva. Per ogni nodo *target*, i corrispondenti *drive* sono individuati monitorando le variazioni nei parametri all'interno di ciascun *Markov blanket* a seguito dell'introduzione nella rete di un'evidenza specifica contenente il massimo valore, cioè 10, per la soddisfazione considerata. L'evidenza si propagherà all'interno della sotto rete attraverso i genitori e i figli del nodo in esame. Per evitare di dilungare oltre il necessario l'esposizione dei risultati, si focalizzerà l'attenzione solo sulle variabili che, caso per caso, subiranno cambiamenti significativi.

SOGIOSET = 10 Inserendo il valore massimo nella soddisfazione dei giorni in cui si effettua la raccolta, la prima distribuzione che riporta dei cambiamenti significativi è la corrispondente importanza. Dalla Figura 3.29 si nota che un aumento di SOGIOSET

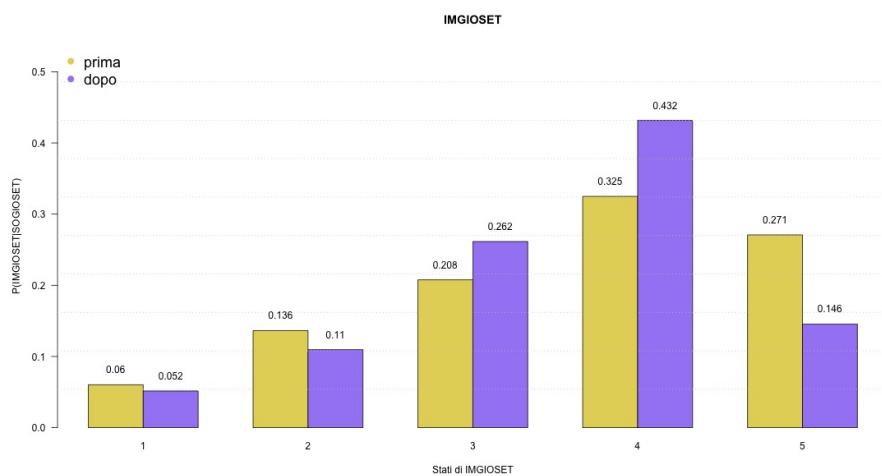


Figura 3.29: Inferenza su SOGIOSET - Importanza giorni raccolta

comporta la diminuzione della probabilità che l'importanza assuma dei valori bassi e il

conseguente aumento della probabilità di voti alti, ad eccezione di '5'.

Anche la soddisfazione sulla disposizione dei container (SODISCON) subisce variazioni

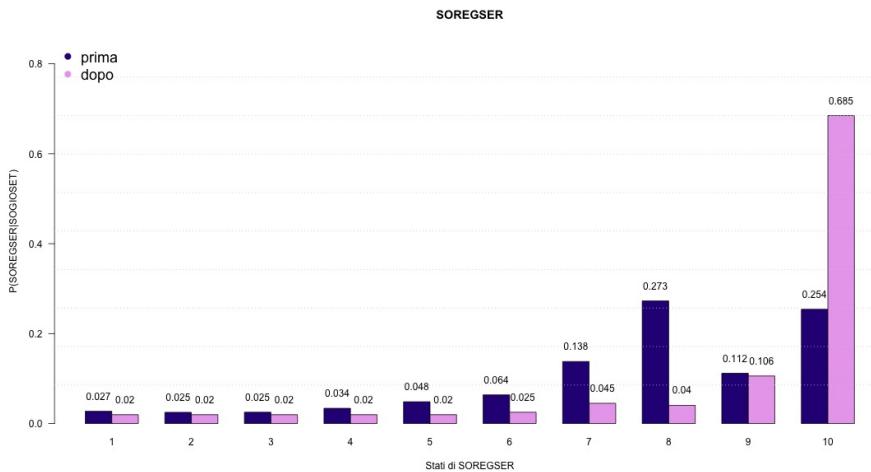


Figura 3.30: Inferenza su SOGIOSET - Regolarità del servizio

rilevanti, il legame però appartiene a quell'insieme di relazioni dubbie che difficilmente possono avere un riscontro nel contesto reale. Discorso analogo anche per la variabile SOCOPRAT, soddisfazione sulla praticità del conferimento nei container.

La relazione più significativa nel *Markov blanket* lega SOGIOSET a SOREGSER, cioè alla soddisfazione sulla regolarità del servizio.

Come risultato dell'inferenza, la distribuzione marginale di SOREGSER (Figura 3.30) si modifica come segue: tutte le probabilità diminuiscono ad eccezione di quella corrispondente al grado di soddisfazione massima, cioè 10, che aumenta più del doppio.

Traducendo in altri termini, sapere che un cittadino è pienamente soddisfatto dei giorni in cui si effettua la raccolta dei rifiuti aumenta la probabilità che dia il voto massimo anche a SOREGSER. Utilizzando il linguaggio delle probabilità si ha che: $P(SOREGSER = 10|SOGIOSET = 10) = 0.685$.

La variabile RIDUUMID (Figura 3.31) fornisce indicazioni utili per aumentare SOGIOSET: gli intervistati maggiormente soddisfatti sono coloro che hanno risposto positivamente all'eventuale riduzione del passaggio dell'umido. Visto che la soddisfazione in esame riguarda il giorni in cui si effettua la raccolta porta a porta, questa relazione ha un possibile riscontro anche nella realtà. Se il comune riuscisse a trovare una soluzione diversa e più accomodante per i cittadini, essi probabilmente risulterebbero più soddisfatti di SOGIOSET.

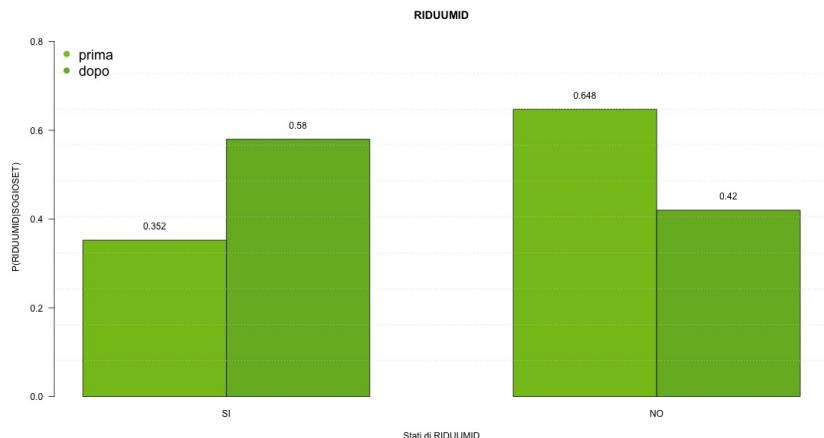


Figura 3.31: Inferenza su SOGIOSET - Riduzione passaggio umido

Le ultime due variabili da analizzare, ad eccezione di quelle anagrafiche come anticipato svolgono solo una funzione descrittiva, sono relative all'orario di conferimento presso i container, SOCOORAR (Figura 3.32a), e le ecostazioni, SOCOECORAR (Figura 3.32b). Anche se appartengono ad aspetti diversi del servizio sono comunque legate tra di loro poiché riguardano il momento in cui si può usufruire del servizio di raccolta rifiuti. Il

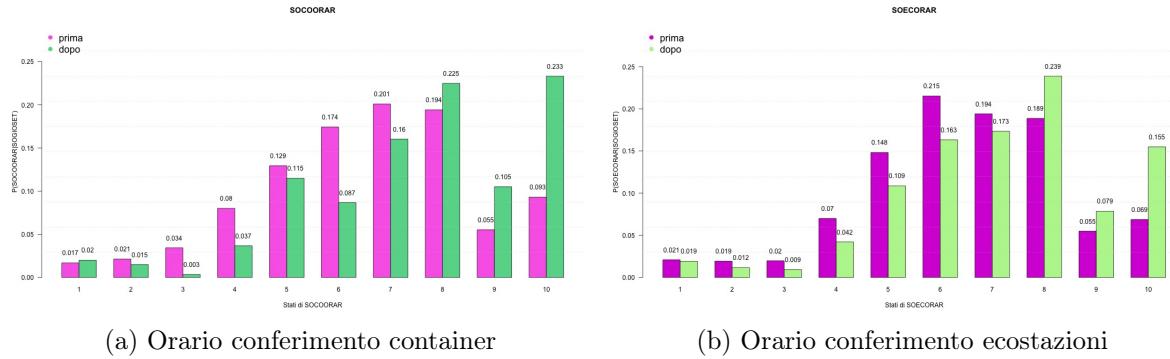


Figura 3.32: Inferenza su SOGIOSET

comportamento delle due variabili dopo l'introduzione dell'evidenza è simile: la probabilità condizionata di tutti gli stati da 1 a 7 diminuisce mentre quella delle valutazioni più alte, cioè da 8 a 10, subisce un incremento. Un cittadino pienamente soddisfatto dei giorni della raccolta ha una maggior probabilità di essere soddisfatto anche degli altri fattori legati agli orari del servizio.

SOREGSER = 10 La prima cosa da valutare è l'effettiva relazione con SOGIOSET: la nuova distribuzione di probabilità (Figura 3.33) mostra che per aumentare la probabilità di ottenere la massima soddisfazione sulla regolarità del servizio bisogna che i cittadini siano molto soddisfatti anche dei giorni in cui si effettua la raccolta. Il ragionamento

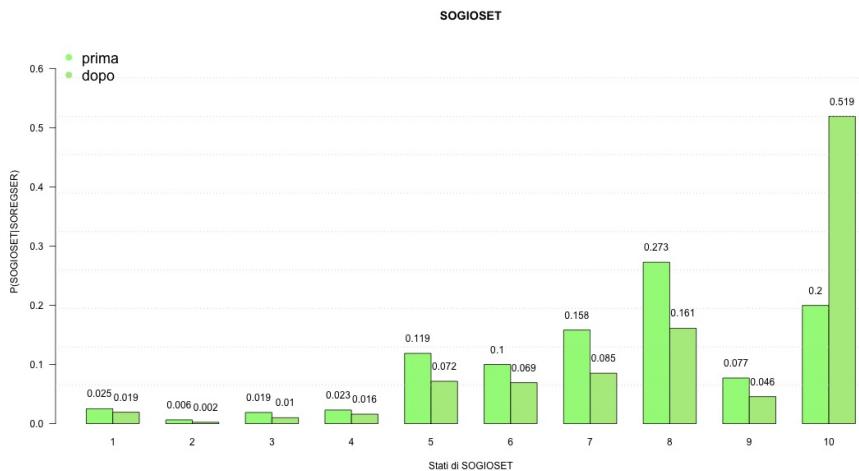


Figura 3.33: Inferenza su SOREGSER - Giorni della settimana

logico alla base di questa considerazione è il seguente: un cittadino potrebbe non essere soddisfatto della regolarità del servizio se il passaggio della raccolta si svolgesse in giornate a lui non ideali. Per massimizzare la soddisfazione sulla regolarità del servizio il

comune dovrebbe cercare di fissare la raccolta porta a porta in giornate che soddisfino il maggior numero di cittadini possibile. Si rilevare la categoria degli individui più insoddisfatti di SOGIOSET con lo scopo di aprire un dialogo con loro e capire come migliorare il servizio.

Massimizzando SOREGSER, si modificano le probabilità degli stati che può assumere la soddisfazione sulla praticità dei container: un cittadino pienamente soddisfatto della regolarità del servizio ha una maggiore probabilità di essere soddisfatto (voti da 7 a 10) anche di SOCOPRAT (Figura 3.34a).

L'ultima variazione significativa avviene all'interno della distribuzione della soddisfazione sul comportamento degli operatori (Figura 3.34b). Introducendo l'evidenza $E \rightarrow (SOREGSER = 10)$, calano le probabilità di tutti gli stati di SOCOMOPE (Figura 3.34b) ad eccezione di quello più alto che registra un elevato incremento: $P(SOCOMOPE = 10|SOREGSER = 10) = 0.587$, segnale di una forte influenza tra queste due soddisfazioni.

I restanti nodi componenti MB di SOREGSER non sono citati poiché le variazioni di ciascuna distribuzione era nulla o trascurabile.

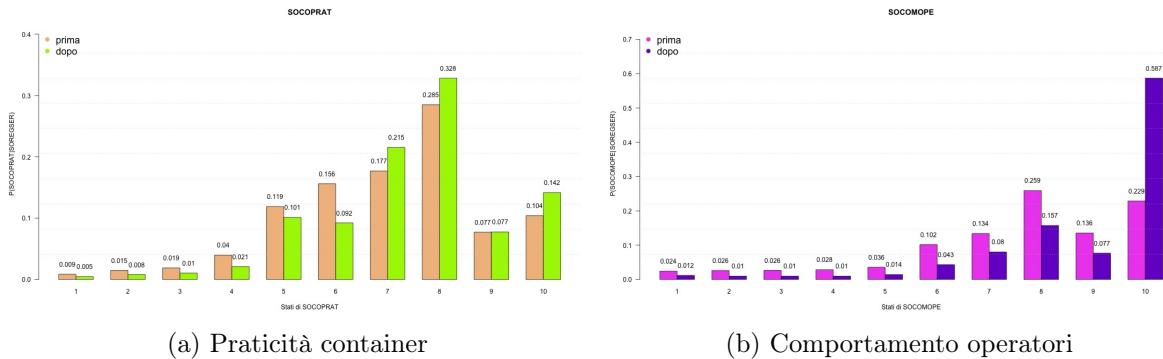


Figura 3.34: Inferenza su SOREGSER

SOCOMOPE = 10 In questo terzo caso, s'introduce nel MB un'evidenza $E \rightarrow (SOCOMOPE = 10)$. Il primo nodo da analizzare è uno dei genitori, PROFOPER. La distribuzione di probabilità marginale (Figura 3.35) di questo nodo si modifica con l'inserimento delle nuove informazioni: le probabilità associate ai voti più bassi (SUFF, SCARSO e INSUFF) diminuiscono mentre aumentano quelle relative ai giudizi positivi (ECC e BUONO).

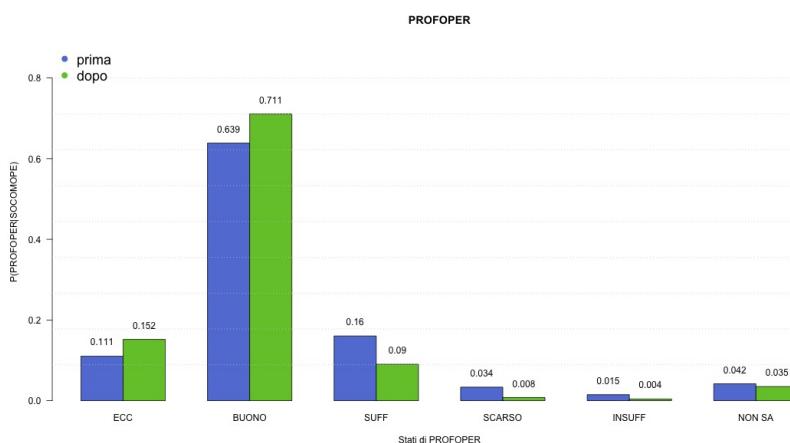


Figura 3.35: Inferenza su SOCOMOPE - Professionalità operatori

Da quanto appena descritto, emerge che migliorando la qualità del lavoro svolto dagli operatori si può migliorare la SOCOMOPE; questo risultato era prevedibile anche senza dover compiere tutta quest'analisi, si ribadisce che nulla è scontato per l'intelligenza artificiale che appren-

de i parametri dai dati.

Il nodo genitore SOREGSER subisce una particolare variazione (Figura 3.36). La probabilità associata a ciascuno stato diminuisce ed aumenta notevolmente la probabilità che la soddisfazione sulla regolarità del servizio sia massima: prima $P(SOREGSER = 10) = 0.291$ mentre ora si ha che $P(SOREGSER = 10|SOCOMOPER = 10) = 0.727$. Questo risultato stabilisce che se il comune vuole migliorare la soddisfazione percepita rispetto al comportamento degli operatori, può essere utile migliorare la regolarità del servizio di raccolta.

Ipotizzando che gli operatori entrino in contatto con i cittadini in prossimità dei container

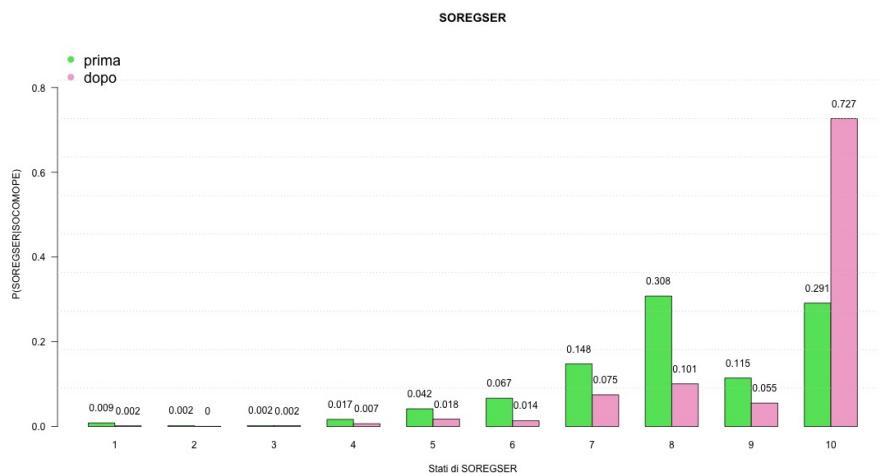
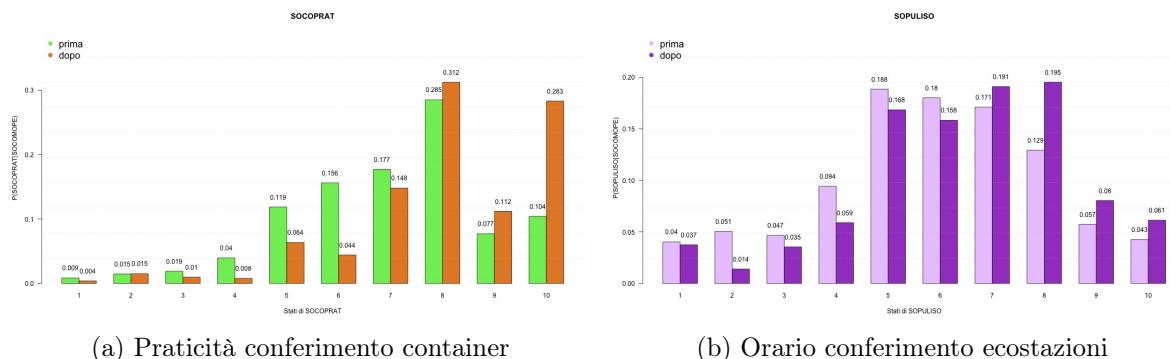


Figura 3.36: Inferenza su SOCOMOPE - Regolarità del servizio

intervenendo per fornire supporto ed aiuto in fase di conferimento⁷, emerge un’importante legame anche con la praticità nei container. In maniera analoga a quanto osservato nei casi precedenti, nella distribuzione di probabilità del nodo SOCOPRAT (Figura 3.37a), sapendo che la soddisfazione sul comportamento degli operatori è massima, aumentano le probabilità associate ai voti più alti, da 8 a 10, e diminuiscono tutti gli altri. Risulta molto probabile che cittadini pienamente contenti del comportamento degli operatori, siano soddisfatti della praticità del conferimento nei container. Un comportamento analogo è



(a) Praticità conferimento container

(b) Orario conferimento ecostazioni

Figura 3.37: Inferenza su SOCOMOPE

tenuto dalla variabile legata alla soddisfazione della pulizia delle isole ecologiche. Se tra

⁷A differenza delle isole ecologiche in cui non ci sono operatori.

le mansioni degli operatori si comprende anche la pulizia del posto di lavoro, allora il loro comportamento è legato a quest'aspetto. La Figura 3.37b mostra come, dopo aver introdotto un alto livello di SOCOMOPE, la probabilità che i cittadini siano soddisfatti positivamente (da 7 a 10) della pulizia delle isole ecologiche aumenta.

SORESSAC = 10 Le relazioni contenute all'interno del *MB* sono alquanto discutibili sul piano pratico, per questo motivo risulta superfluo descrivere i risultati ottenuti.

SODISISO = 10 Le informazioni più interessanti si riscontrano nelle variazioni che subiscono i nodi genitori di SODISISO. Due di essi riguardano la soddisfazione sulla disposizione di altri due aspetti della raccolta dei rifiuti, i container e le ecostazioni. Questo sta a significare che le collocazioni sul territorio comunale di questi servizi sono tra loro collegate. Sapere che il cittadino è completamente soddisfatto della disposizione delle

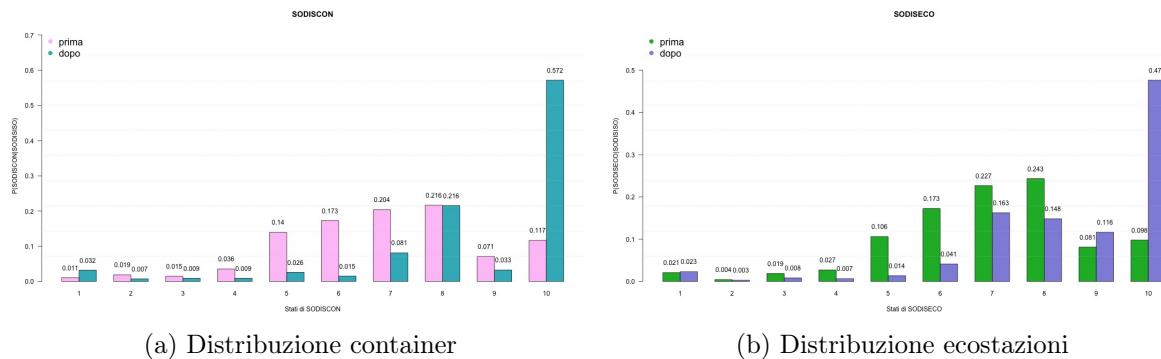


Figura 3.38: Inferenza su SODISISO

isole ecologiche aumenta di circa 5 volte la probabilità che lo sia anche della distribuzione (Figura 3.38a) dei container; le probabilità associate agli altri stati invece diminuiscono. Queste conclusioni sono valide anche per la distribuzione delle ecostazioni, anche se con altri valori assoluti (Figura 3.38b). Da quest'analisi si conclude che se il comune volesse aumentare SODISISO, migliorare la soddisfazione sulla distribuzione di tutti gli elementi del servizio di raccolta rifiuti sarebbe probabilmente efficace.

Le probabilità marginali dell'ultimo genitore, cioè PROFOPER, non subiscono variazioni significative dall'introduzione di un'alta soddisfazione di SODISISO nella rete.

Le altre due variabili appartenenti al gruppo delle 'isole ecologiche' sono la soddisfazione sulla capienza delle campane e sulla pulizia di queste aree. Il primo di questi elementi, la capienza dei cassonetti (Figura 3.39a), ha un comportamento insolito: il lato positivo è l'aumento della probabilità marginale associata agli stati 8 e 10 insieme alla contestuale diminuzione di quelle relative ai valori più bassi, la particolarità è data dall'aumento della probabilità del valore 5 e dalla riduzione dello stato 9. La pulizia delle isole ecologiche (Figura 3.39b) presenta variazioni simili a quelle osservate in precedenza: chi è soddisfatto della distribuzione delle isole ecologiche è probabilmente disposto a valutare positivamente anche il modo in cui gli operatori le tengono pulite.

SOCAPCAM = 10 L'evidenza introdotta nel *MB* si propaga con le seguenti conseguenze. Gli individui che si reputano soddisfatti al massimo di SOCAPCAM, tuttavia, non attribuiscono molta importanza a quest'aspetto.

La Figura 3.40 mostra che per il comune non è necessario sensibilizzare i cittadini sulla

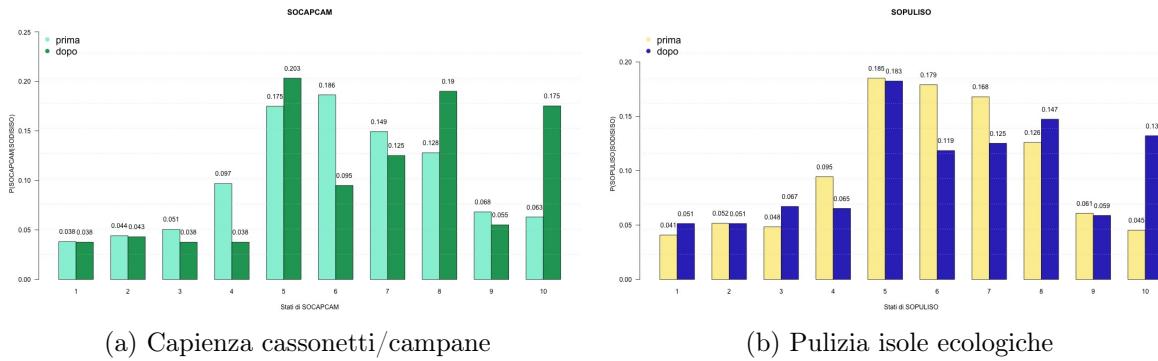


Figura 3.39: Inferenza su SODISISO

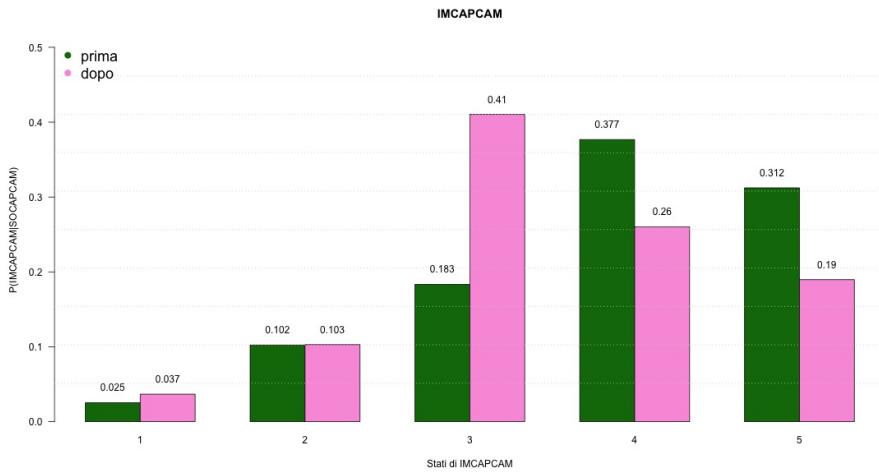


Figura 3.40: Inferenza su SOCAPCAM - Importanza capienza cassonetti

dimensione delle campane presenti nelle isole ecologiche perché questi ne risultino soddisfatti. IMCAPCAM non è, perciò, uno dei *driver* di questa soddisfazione.

La capienza dei cassonetti è traducibile come indice di praticità del conferimento, il che renderebbe plausibile il collegamento con SOCOPRAT. In particolare, la relazione tra i due nodi è tale che un ottimo livello di SOCAPCAM implica un aumento della probabilità che la variabile SOCOPRAT assuma dei valori alti, da 8 a 10.

Il comportamento insolito della relazione tra la capienza dei cassonetti e la distribuzione delle isole ecologiche è confermata anche in questa situazione. Il probabile legame tra i due aspetti è al diminuire della capienza della campane nelle isole ecologiche per mantenere lo stesso livello di soddisfazione è necessario che aumenti la loro

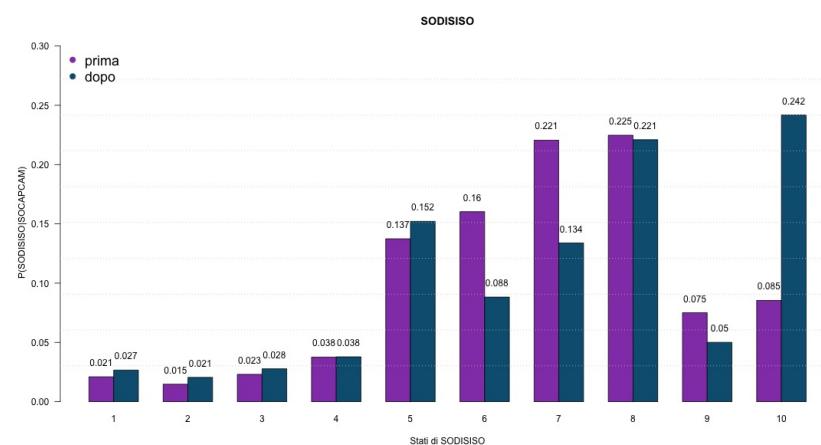


Figura 3.41: Inferenza su SOCAPCAM - Distribuzione isole ecologiche

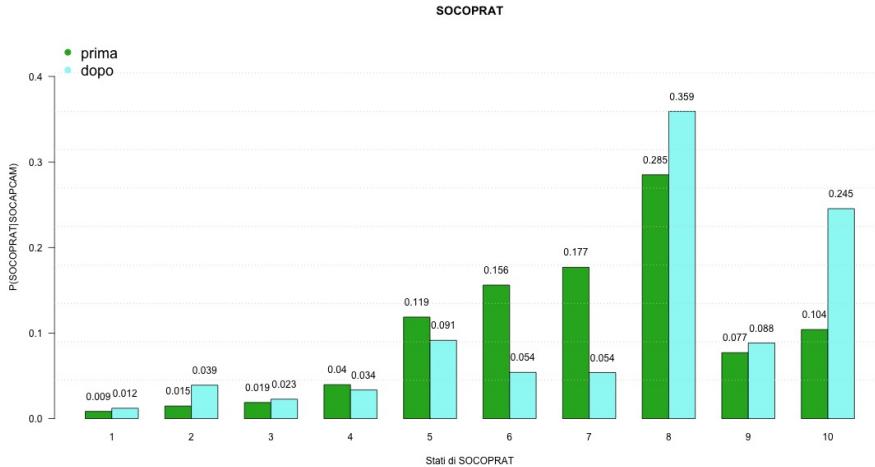


Figura 3.42: Inferenza su SOCAPCAM - Praticità conferimento nei container

distribuzione sul territorio comunale. Il motivo è che nel caso in cui le campane siano piene e non sia possibile conferire rifiuti in un’isola ecologica, i cittadini sono costretti a raggiungere un’altra di queste aree. La Figura 3.41 mostra che indubbiamente, un alto indice di soddisfazione sulla capienza delle campane, in virtù di quanto appena detto, comporti un aumento della probabilità che anche SODISISO assuma un valore elevato. Tuttavia, si conferma la diminuzione della $P(SODISISO = 9|SOCAPCAM = 10)$ e il contestuale aumento, se pur minimo, dei valori più bassi (da 1 a 5).

Le variazioni che subisce la distribuzione di probabilità marginale di SOPULISO, anche se di entità minima, sono analoghe a quelle appena descritte per SODISISO e sono riportate nella Figura 3.43. Si può notare un leggero aumento della probabilità dei valori

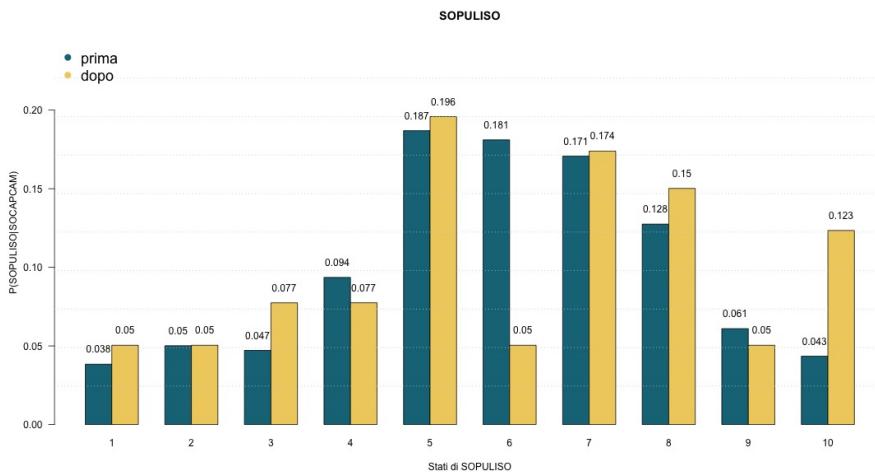


Figura 3.43: Inferenza su SOCAPCAM - Pulizia isole ecologiche

indicanti giudizi sia negativi, da 1 a 5, sia positivi, cioè 7, 8, 10; in particolare la soddisfazione massima triplica la propria probabilità a seguito della propagazione dell’evidenza. Risulta difficile interpretare la diminuzione legata agli stati 6 e 9.

SOPULISO = 10 La prima cosa da segnalare è la presenza di un legame difficilmente giustificabile in un contesto reale, quello tra SOPULISO e SOECORAR. I dubbi su

questo legame sono confermati dal fatto che non due nodi vicini, inoltre la distribuzione di probabilità sulla soddisfazione dell'orario di conferimento nelle ecostazioni subisce variazioni trascurabili a seguito dell'introduzione dell'evidenza. Nonostante ci siano dei cambiamenti di lieve entità nella probabilità marginale, anche l'arco tra SOPULISO e SOCOPRAT difficilmente restituisce utili informazioni da calare nel contesto reale. Un aspetto interessante è che le distribuzioni di probabilità dei tre nodi genitori, SOCOMOPE, SOCAPCAM e SODISISO, abbiano lo stesso andamento, differiscono solo per i valori associati alle probabilità dei diversi stati. In linea di massima, le nuove informazio-

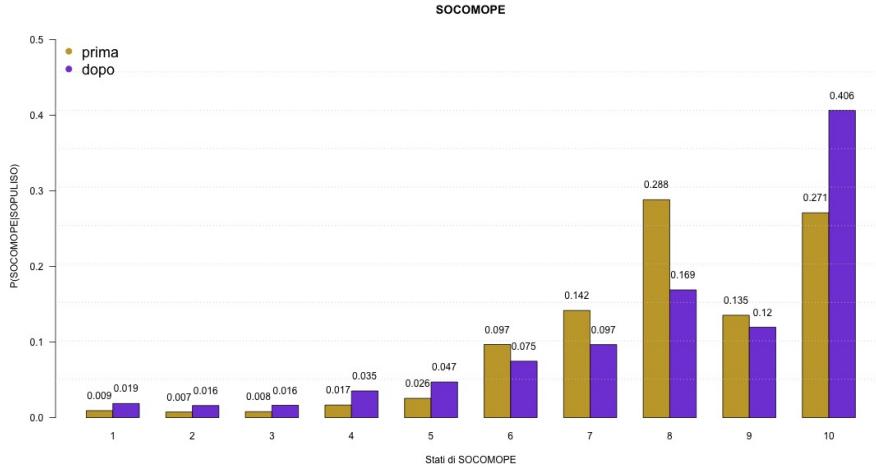


Figura 3.44: Inferenza su SOPULISO - Comportamento operatori

ni si propagano nella rete diminuendo le probabilità degli stati che si collocano nel mezzo dalla distribuzione (da 4 a 8) e aumentando la quella delle code (stati 1,2,3,9,10). Nonostante l'incremento della loro probabilità, gli stati connessi a segnali d'insoddisfazione non raggiungono valori assoluti sufficienti da mettere il comune in allerta, poiché in questo caso, bisognerebbe intraprendere una strategie che eviti di aumentare la soddisfazione sulla pulizia delle isole ecologiche perché vorrebbe dire incoraggiare la probabilità che le insoddisfazioni connesse ad altri tre fattori aumentino. I grafici raffiguranti le distribuzio-

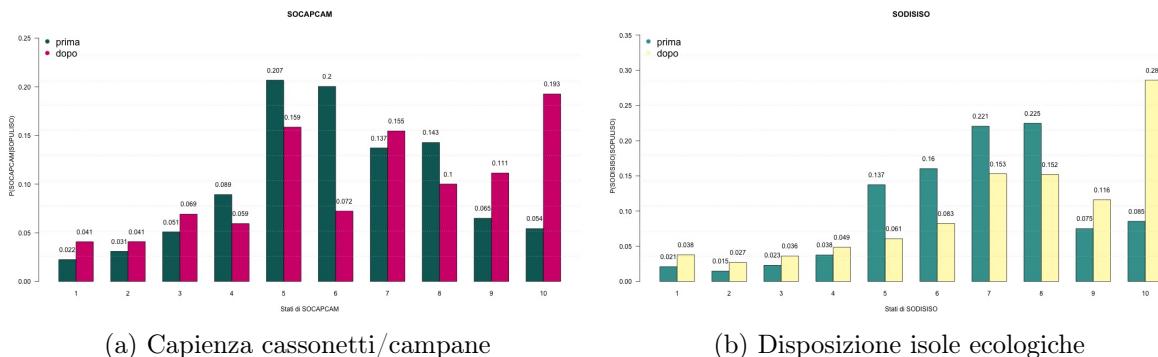


Figura 3.45: Inferenza su SOPULISO

ni di probabilità (Figure 3.44, 3.45a e 3.45b) contengono un segnale positivo: la variazione più significativa è quella della massima soddisfazione, cioè il 10. Attuando una strategia focalizzata sulla soddisfazione della pulizia delle isole ecologiche il comune può aspettarsi un incremento della probabilità che i cittadini siano pienamente soddisfatti della capienza

delle campane, del comportamento degli operatori e della disposizione delle isole ecologiche sul territorio.

L'ultima variabile da analizzare rappresenta un indice di pulizia ed è la soddisfazione della frequenza con cui sono pulite le strade. Di tre cittadini pienamente soddisfatti della pulizia delle isole ecologiche, uno è altrettanto soddisfatto anche della frequenza di spazzamento delle strade, infatti: $P(SOFRESER = 10 | SOPULISO = 10) = 0.254$, valore tre volte tanto rispetto a prima dell'evidenza. Si segnala che la probabilità degli stati

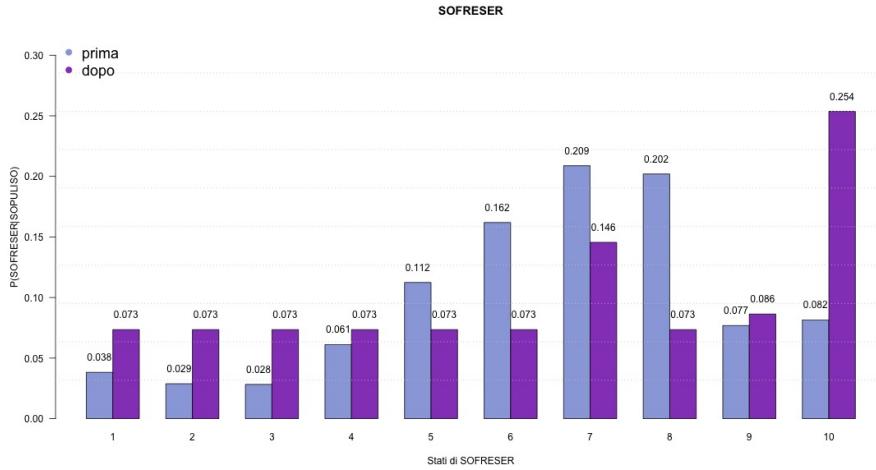


Figura 3.46: Inferenza su SOPULISO - Frequenza spazzamento strade

1,2,3,4,5,6 ed 8 dopo la propagazione delle nuove informazioni è uguale.

SODISCON = 10 Ponendo la soddisfazione sulla distribuzione dei container al massimo, si ottengono i seguenti risultati. I parametri della variabile relativa alla soddisfazione sull'orario di conferimento nei container variano significativamente (Figura 3.47) in modo simile alle precedenti: i valori relativi agli stati più alti, cioè 9 e 10, aumentano la propria probabilità di essere attribuiti dal cittadino a quest'aspetto del servizio mentre calano le probabilità di tutti gli altri stati.

L'attenzione è ora rivolta alla distribuzione degli altri aspetti del servizio. Per quanto riguarda la SODISISO (Figura 3.48a) e SODISECO (Figura 3.48b), i cambiamenti registrati nelle probabilità marginali dopo l'introduzione dell'evidenza sono gli stessi appena descritti per SOCOORAR: un cittadino che attribuisca voto 10 alla distribuzione dei container, sarà probabilmente più soddisfatto, voti da 8 a 10, rispetto al resto del campione anche della distribuzione delle isole ecologiche e delle ecostazioni.

Descrivendo i *MB* di ogni variabile, è emerso un collegamento tra le ecostazioni e i container. Il propagarsi dell'evidenza all'interno della sotto rete porta alla variazione della distribuzione di probabilità delle due variabili legate alle ecostazioni ancora da analizzare: la praticità di conferimento nelle ecostazioni (Figura 3.49a) e l'orario di conferimento (Figura 3.49b). In entrambi i casi, cresce la probabilità a posteriori degli stati 9 e 10 e diminuisce quella di tutti gli altri valori.

Tra i genitori di SODISCON c'è il pensiero degli intervistati sulla perdita di decoro del centro storico. Prima dell'evidenza, tutti gli stati della variabile, ad eccezione di "Non sa", sono equiprobabili; successivamente la distribuzione cambia ed aumentano le probabilità sia delle risposte positive, "No" e "Poco", sia di quella negativa "Molto". In questa situazione, dato che un aumento di SODISCON si può ottenere anche se DECCENST è

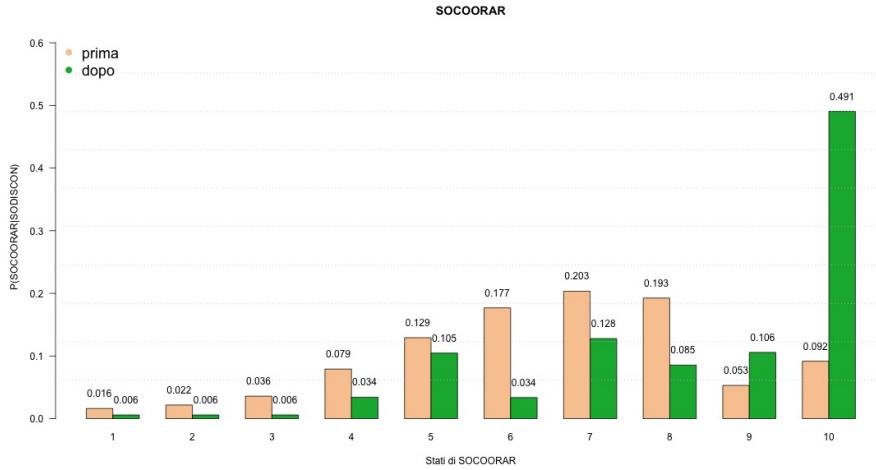


Figura 3.47: Inferenza su SODISCON - Orario conferimento container

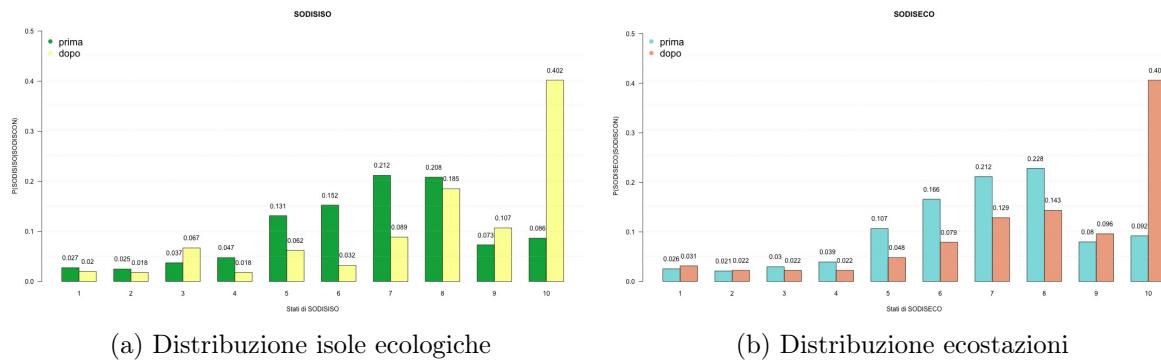


Figura 3.48: Inferenza su SODISCON

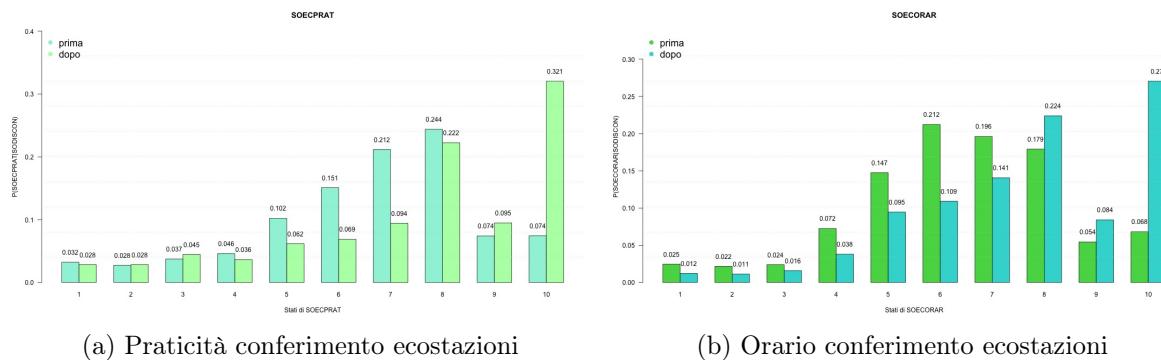


Figura 3.49: Inferenza su SODISCON

negativa, potrebbe essere utile dare un’occhiata anche al *MB* del nodo genitore in modo da determinare i fattori sui quali agire per incrementare il numero di giudizi positivi sul decoro del centro storico, incremento che a sua volta porterebbe ad una maggior soddisfazione della distribuzione delle isole ecologiche.

Il secondo nodo genitore è la soddisfazione sui giorni in cui si effettua la raccolta. La nuova distribuzione di probabilità (Figura 3.51a) mostra che chi è completamente soddisfatto della distribuzione dei container ha una probabilità molto alta, $P(SOGIOSET = 10|SODISCON = 10) = 0.607$, di essere pienamente soddisfatto anche dei giorni in cui

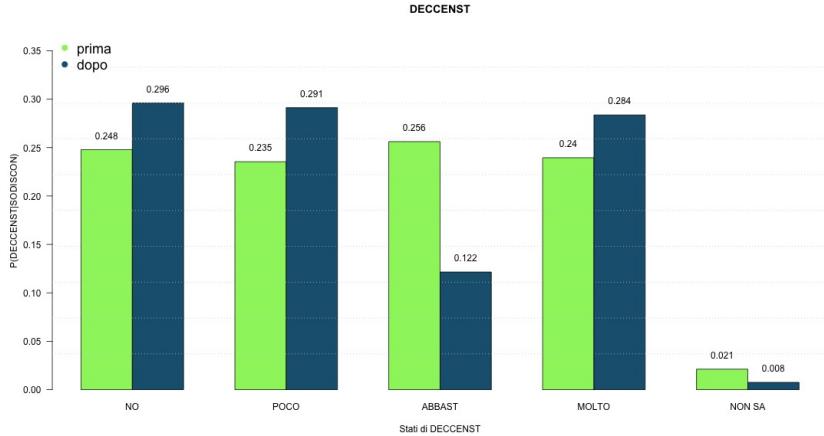


Figura 3.50: Inferenza su SODISCON - Decoro centro storico

si effettua la raccolta. Questa relazione fa sorgere qualche perplessità su un’eventuale applicazione nel contesto reale.

Discorso analogo anche per SORESSAC (Figura 3.51b), introducendo l’evidenza $E \rightarrow SODISCON = 10$ nel modello, aumenta la probabilità che la soddisfazione sulla resistenza dei sacchetti assuma gli stati da 7 a 10, indicatori di un alto gradimento. Questa relazione statistica individuata dall’intelligenza artificiale, degli approfondimenti prima di essere applicata.

Dopo aver descritto come si propagano le informazioni all’interno di questa sotto rete, si

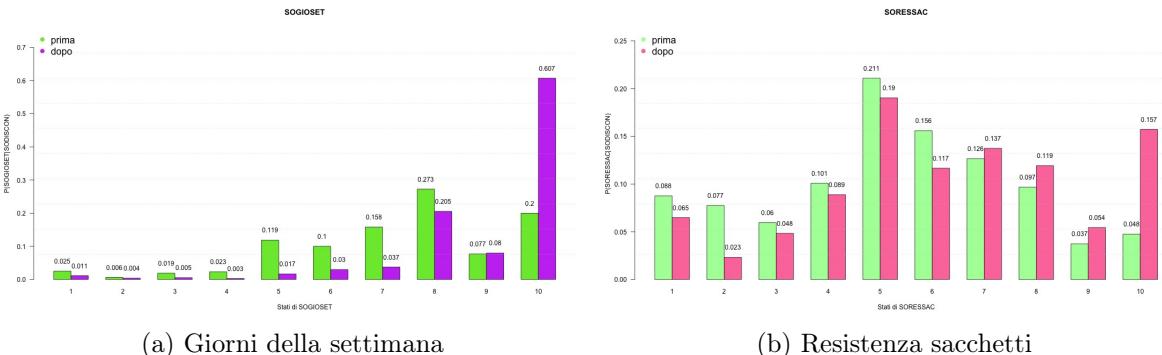


Figura 3.51: Inferenza su SODISCON

può concludere che il comune può beneficiare delle sinergie individuate: aumentando la soddisfazione riguardante la distribuzione delle isole ecologiche è possibile migliorare la valutazione attribuita dai cittadini alle ecostazioni e alla distribuzione degli altri elementi del servizio.

SOCORAR = 10 In prima battuta si analizzano le due variabili relative alla soddisfazione sui container. Le distribuzioni marginali degli stati dei due nodi sono molto simili, calano le probabilità di tutti i giudizi sulla soddisfazione ad eccezione di 10 che aumenta di circa sei volte in entrambi i casi. Questo significa che su 10 individui che valutino 10 la propria SOCORAR, circa 6 attribuirebbero il medesimo punteggio anche alla SODISCON (Figura 3.52a) o alla SOCOPRAT (Figura 3.52b). Questi risultati suggeriscono che per migliorare la soddisfazione sugli orari di conferimento, il comune

dovrebbe aumentare il grado di soddisfazione relativo agli elementi dei container. Il ra-

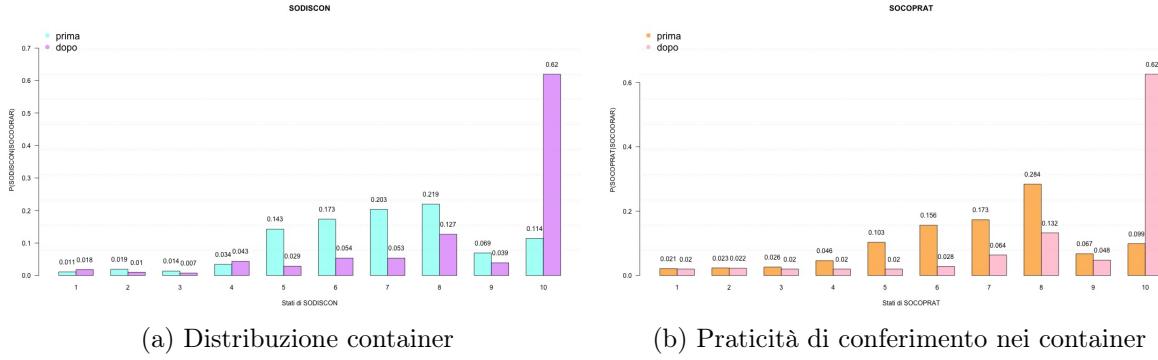


Figura 3.52: Inferenza su SOCOORAR

gionamenti alla base di questa considerazione sono probabilmente i seguenti: in primo luogo, se i container sono ben distribuiti sul territorio, il cittadino ha minori difficoltà a raggiungerli, perciò non avrà particolari esigenze per quanto riguarda gli orari in cui può accedervi; in secondo luogo, maggiore è la semplicità e praticità del conferimento e minore sarà il tempo impiegato per il conferimento, per cui non sono richiesti orari particolari. La Figura 3.53 mostra la distribuzione di probabilità di IMDISCON prima e dopo l'introduzione dell'evidenza nella rete, ciò che si evince è che la metà dei cittadini soddisfatti pienamente dell'orario di conferimento nei container considera molto importante anche la loro distribuzione sul territorio. Il risultato ottenuto conferma quanto detto sull'influenza tra distribuzione e orario dei container.

La variabile anagrafica relativa al titolo di studi (Figura 3.54) informa il comune su qua-

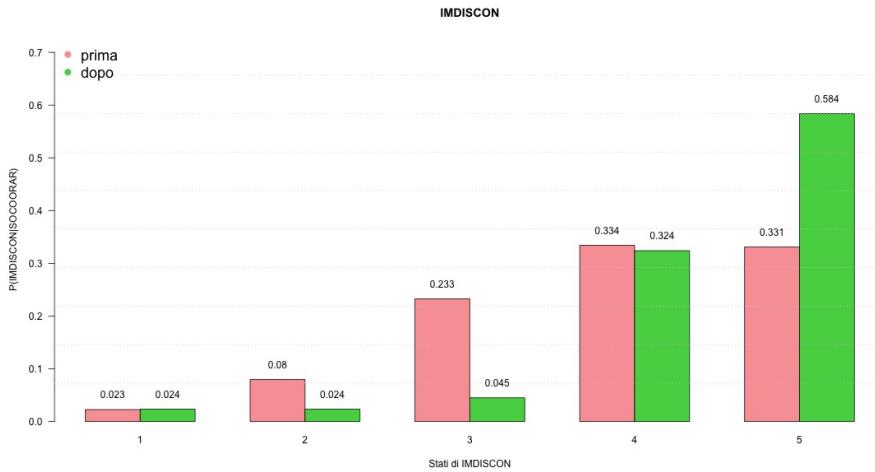


Figura 3.53: Inferenza su SOCOORAR - Distribuzione container - Importanza

le sia la probabile formazione scolastica dei cittadini che abbiano dato il voto massimo a SOCOORAR: la probabilità più alta è $P(TITSTUD = OBBLIGO|SOCOORAR = 10) = 0.666$ ed appartiene a coloro che hanno frequentato solo la scuola dell'obbligo mentre i laureati sono la categoria meno probabile, quasi impossibile.

In ultima battuta, si considerano due variabili che, come SOCOORAR, esprimono la soddisfazione sugli orari legati ai servizi di raccolta rifiuti, SOGIOSET e SOECORAR. In entrambe le distribuzioni di probabilità marginale (Figure 3.55a e 3.55b), lo stato più

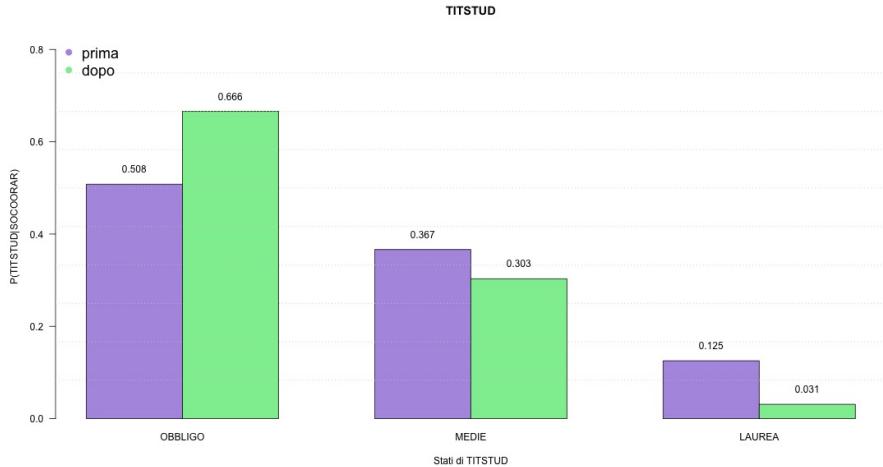


Figura 3.54: Inferenza su SOCOORAR - Titolo di studio

probabile è il 10. Anche in questo caso, la massima soddisfazione sugli orari dei container comporta una maggior probabilità che i cittadini siano contenti anche degli orari di conferimento nelle ecostazioni e dei giorni in cui si effettua la raccolta porta a porta.

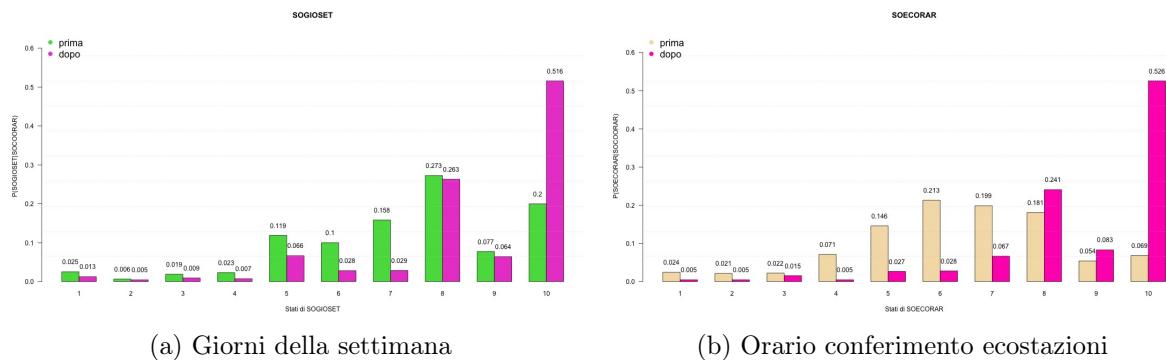


Figura 3.55: Inferenza su SOCOORAR

SOCOPRAT = 10 Uno dei genitori di SOCOPRAT è il numero di componenti del nucleo familiare. La distribuzione (Figura 3.56) di questa variabile a seguito dell'evidenza, informa il comune che gli individui soddisfatti della praticità di conferimento nei container sono probabilmente coloro che vivono in coppia. Al contrario, gli intervistati che vivono in nuclei numerosi sono probabilmente meno soddisfatti di quest'aspetto. Attraverso indagini più specifiche, il comune potrebbe individuare le cause legate all'insoddisfazione per poi determinare una strategie efficacie per migliorare il pensiero dei cittadini.

Le variazioni dei parametri di SOFRESER non sono riportate poiché la relazione rientra nel gruppo di legami 'dubbi', da approfondire prima che siano utilizzati per giungere a conclusioni pratiche.

Per quanto concerne gli altri aspetti dei container, solo il nodo relativo agli orari è presente nella sotto rete e fa parte dei genitori di SOCOPRAT. La Figura 3.57 mostra la distribuzione di probabilità del nodo prima e dopo l'introduzione dell'evidenza nel modello. Il risultato ottenuto è un aumento della probabilità che un cittadino soddisfatto

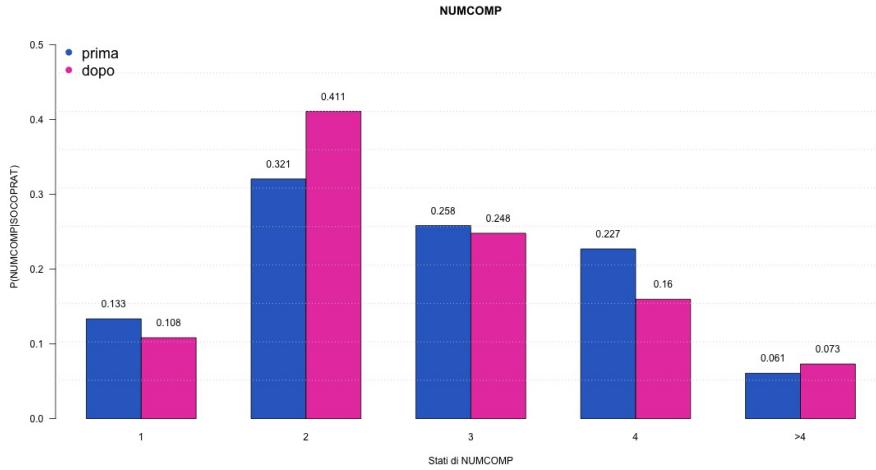


Figura 3.56: Inferenza su SOCOPRAT - Numero di componenti

della praticità di conferimento nei container, sia altrettanto soddisfatto della praticità del conferimento in queste aree. Utilizzando il linguaggio delle probabilità si dice che: $P(SOCOORAR = 10 | SOCOPRAT = 10) = 0.458$. Richiamando l’ipotesi compiuta

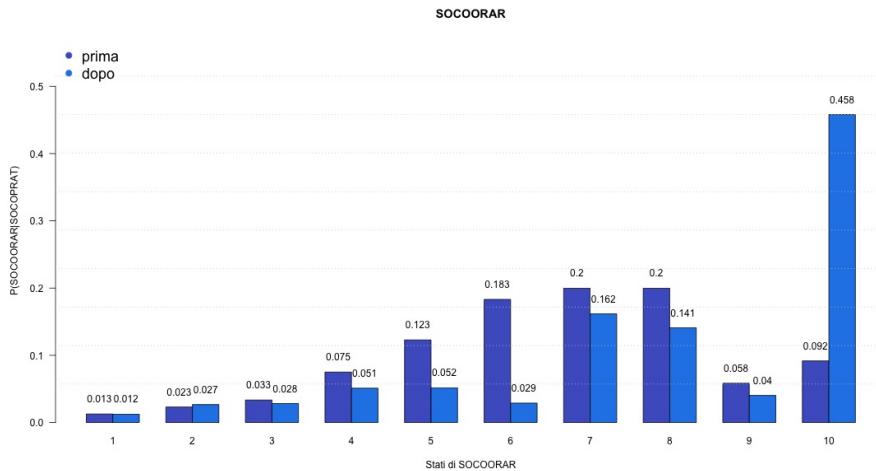


Figura 3.57: Inferenza su SOCOPRAT - Orario conferimento container

in precedenza per cui SOCAPCAM è un indice di praticità del conferimento nelle isole ecologiche, non destava alcuna sorpresa osservare che l’evidenza ha causato delle variazioni nella sua distribuzione di probabilità marginale (Figura 3.58a). Dopo la propagazione delle informazioni è aumentata la probabilità che anche SOCAPCAM riceva valutazioni molto positiva, da 8 a 10. Lo stesso discorso vale anche per l’ultimo indice di praticità che è quello nelle ecostazioni, SOECPRAT (Figura 3.58b). Questo significa che manipolando SOCOPRAT il comune può agire positivamente anche gli altri aspetti legati alla praticità della raccolta rifiuti.

Ragionevolmente si può ipotizzare che nelle aree dedicate ai container ci siano degli operatori che aiutano i cittadini nel conferimento. Per questo motivo il nodo figlio SOCOMOPE subisce l’influenza dell’evidenza su SOCOPRAT. Nello specifico, nella Figura 3.59 si assiste ad una riduzione della probabilità di tutti gli stati ed eccezione del punteggio massimo che passa da $P(SOCOMOPE = 10) = 0.223$ a $P(SOCOMOPE = 10 | SOCOPRAT =$

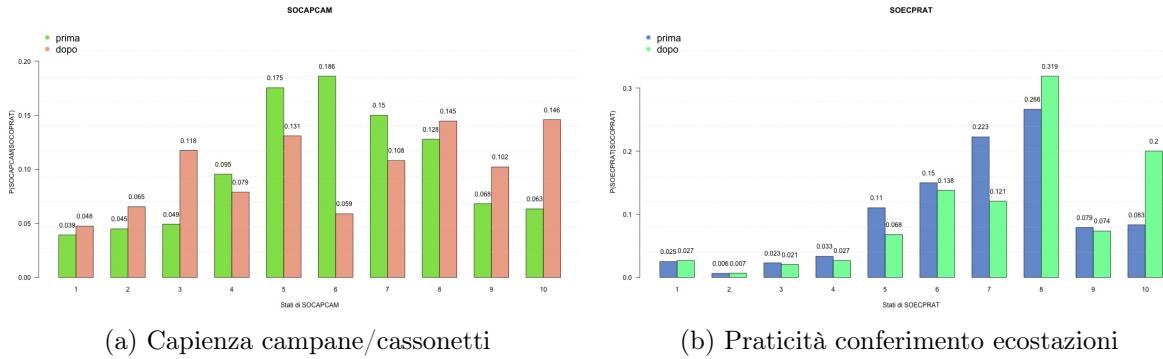


Figura 3.58: Inferenza su SOCOPRAT

10) = 0.54. Quest'aumento considerevole porta a concludere che un incremento della soddisfazione nella praticità di conferimento è probabilmente dato da un miglioramento nel comportamento degli operatori.

Si aggiungono alla lista delle relazioni che necessitano di maggiori approfondimenti quelli

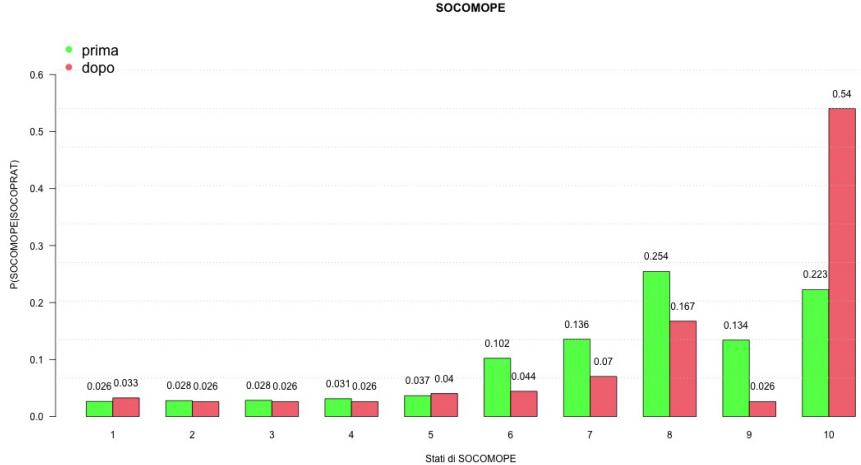


Figura 3.59: Inferenza su SOCOPRAT - Comportamento operatori

che coinvolgono i nodi SOPULISO e SOGIOSET.

SODISECO = 10 I vicini di questo nodo descrivono la soddisfazione o della distribuzione degli altri elementi del servizio o delle altre caratteristiche delle ecostazioni. La prima coppia di nodi, rispettivamente SODISISO (Figura 3.60a) e SODISCON (Figura 3.60b) subiscono variazioni molto simili a seguito della propagazione delle nuove informazioni; diminuisce considerevolmente la probabilità associata a tutti gli stati minori o uguali a 7 mentre aumenta quella dei valori di soddisfazione più elevati, da 8 a 10. Le strategie che mirano ad aumentare la soddisfazione sulla distribuzione delle ecostazioni possono contenere delle sinergie che portano ad un aumento anche di SODISISO e SODISCON.

Il gruppo di nodi che descrive le caratteristiche delle ecostazioni è caratterizzato da forti legami d'influenza al suo interno. Introducendo nel modello un'evidenza forte del tipo *SODISECO* = 10, i nodi sulla soddisfazione dell'orario delle ecostazioni (Figura 3.61a) e sulla praticità del conferimento (Figura 3.61b) subiscono dei miglioramenti che si tra-

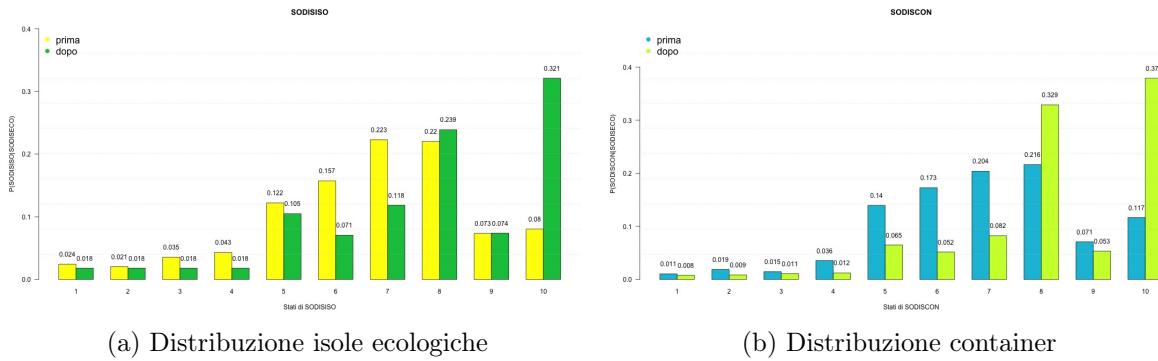


Figura 3.60: Inferenza su SODISECO

ducono in un aumento della probabilità di giudizi di piena soddisfazione, 9 e 10 e la contestuale diminuzione delle valutazioni più basse, da 1 a 7. Migliorando la distribu-

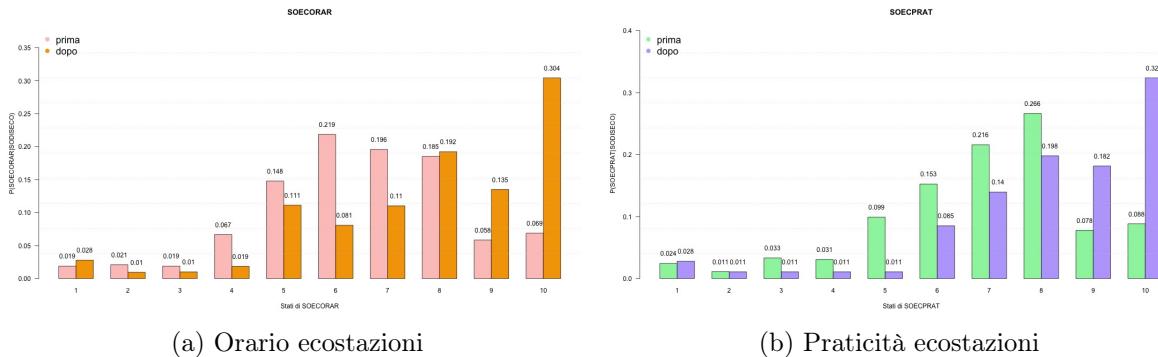


Figura 3.61: Inferenza su SODISECO

zione delle ecostazioni sul territorio comunale il comune può beneficiare di un probabile aumento del grado di soddisfazione di tutti gli altri aspetti legati alle ecostazioni. Il legame tra praticità distribuzione e orari è lo stesso ipotizzato per i container. La distribuzione di probabilità marginale della professionalità degli operatori non subisce variazioni significative dove "Buono" rimane la valutazione più probabile.

SOECORAR = 10 La variabile ATTATTU, indicante l'attività lavorativa, non subisce particolari variazioni; la classe più probabile rimane quella dei "pensionati" seguiti dai "dipendenti", ultimi invece i "disoccupati". Conoscere il giudizio di un individuo sulla soddisfazione dell'orario nelle ecostazioni non porta novità nelle probabilità legate alla sua professione. Lo stesso discorso vale per il nodo GENERE, PROFOPER, SOFRESER, IMREGSER, IMDISCON, IMGIOSET e IMCOPRAT. A differenza di altre variabili come QUARTIER che mantengono completamente immutata la propria distribuzione di probabilità marginale anche dopo l'evidenza, i nodi appena richiamati subiscono delle variazioni ma di valore assoluto troppo basso per essere prese in considerazione. A titolo esemplificativo si guardi la Figura 3.62.

Il nodo in esame esercita una forte influenza su uno dei suoi genitori, SOCOORAR, segno che migliorando profondamente la soddisfazione sull'orario delle ecostazioni il comune può beneficiare di un probabile miglioramento anche di quello dei container. La Figura 3.63 indica che l'evidenza introdotta nel modello riduce la probabilità di tutti gli stati di SOCOORAR ad eccezione di quello maggiore che addirittura raggiunge un punteggio

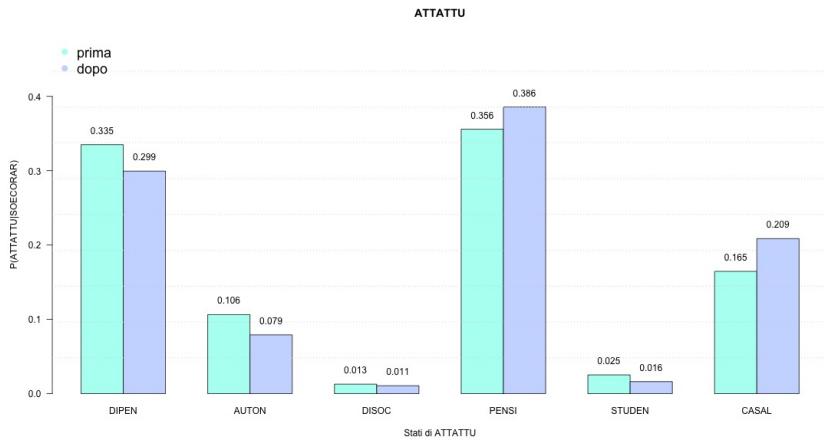


Figura 3.62: Inferenza su SOECORAR - Attività lavorativa

di $P(SOCOORAR = 10|SOECORAR = 10) = 0.702$. Significa che c'è una probabilità del 70% che un cittadino soddisfatto al massimo di SOECORAR lo sia altrettanto anche dell'orario dei container.

Ulteriori variabili che trattano degli orari, in senso lato, dei servizi di raccolta dei rifiuti

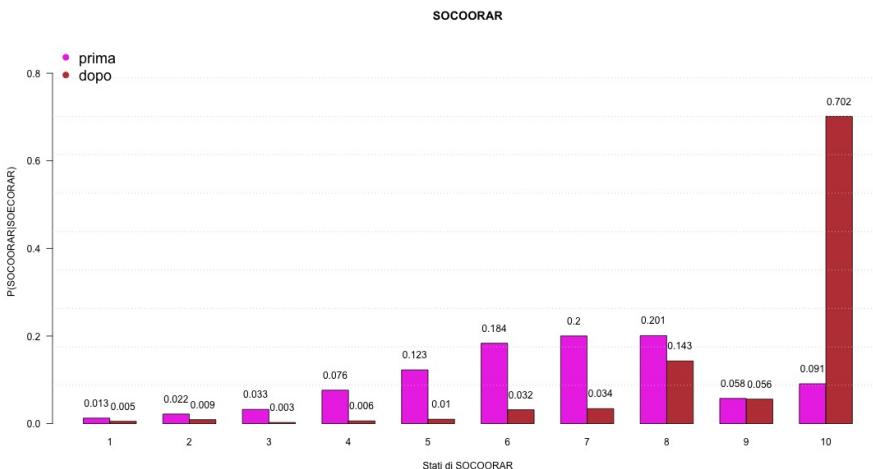


Figura 3.63: Inferenza su SOECORAR - Orario di conferimento container

sono la soddisfazione sui giorni della raccolta porta a porta e la conseguente regolarità del servizio. Le modifiche nelle distribuzioni (Figure 3.64a e 3.64b) di probabilità sono simili a quelle descritte per SOCOORAR: il livello massimo di soddisfazione diventa lo stato più probabile a discapito di tutti gli altri valori. Quanto appena scoperto conferma i legami tra le variabili legate agli orari dei diversi servizi offerti.

In ultima battuta si considerano i due aspetti residuali delle ecostazioni. Dal confronto tra le nuove e le precedenti distribuzioni di probabilità emerge quanto segue: sapere che un cittadino è pienamente soddisfatto degli orari di conferimento presso le ecostazioni aumenta notevolmente la probabilità che egli sia altrettanto soddisfatto anche della distribuzione (Figura 3.65a) e della praticità di conferimento (Figura 3.65b) presso le ecostazioni. Una strategia che agisce sul miglioramento di SOECORAR avrà un ritorno positivo anche sulla soddisfazione di tutti gli aspetti delle ecostazioni.

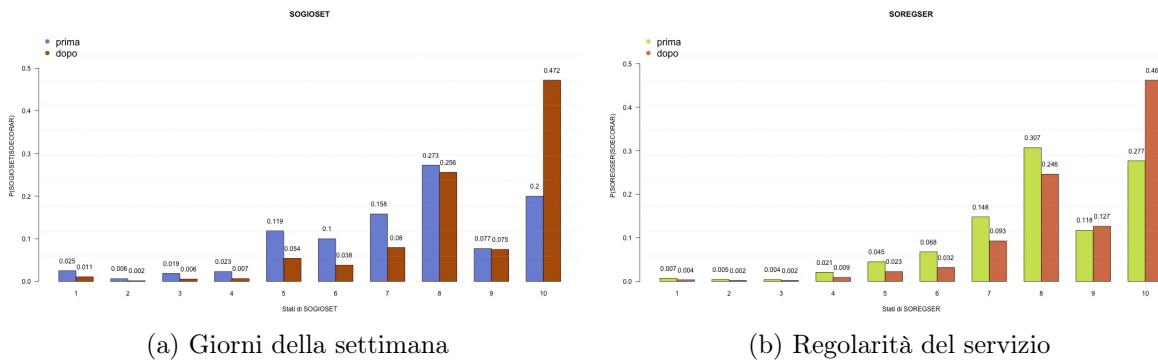


Figura 3.64: Inferenza su SOECORAR

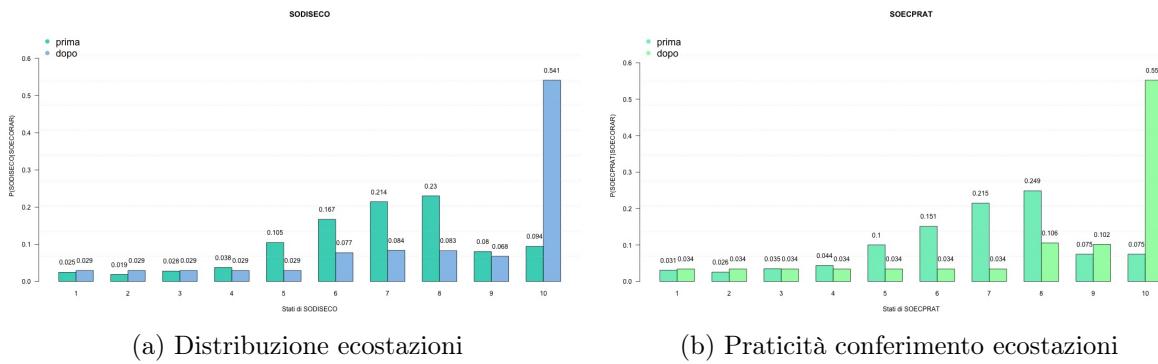


Figura 3.65: Inferenza su SOECORAR

Le variazioni subite dai nodi IMCOMOPE, SODISCES, SODISCON non sono descritte poiché fondate su relazioni di dubbia applicabilità pratica.

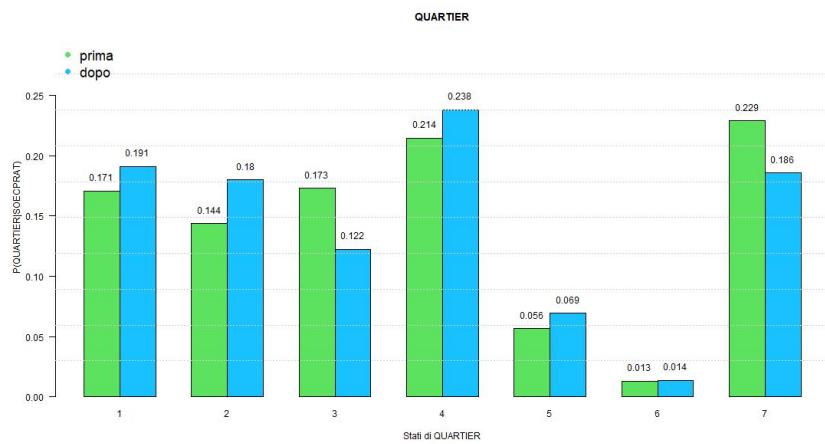
SOECPRAT = 10 Il nodo che descrive il quartiere di residenza degli intervistati non subisce variazioni significative nella propria distribuzione di probabilità: i cittadini più soddisfatti hanno una maggior probabilità di risiedere nel quarto quartiere mentre il sesto rimane quello meno probabile di tutti.

Chi abita in quest'ultimo quartiere difficilmente è soddisfatto pienamente della praticità delle ecostazioni, ulteriori approfondimenti potrebbero far emergere le cause di quest'insoddisfazione. Osservando la distribuzione (Figura 3.66), si può concludere che probabilmente, l'ecostazione di cui si avvolgono abitualmente i residenti

Figura 3.66: Inferenza su SOECPRAT - Quartiere di residenza

nel quinto e sesto quartiere presenta delle difficoltà di conferimento che creano malcontento tra i cittadini.

Il propagarsi dell'evidenza nella sotto rete ha portato a delle variazioni nei parametri



legati alle altre variabili delle ecostazioni. In particolare li soddisfazioni sia della distribuzione (Figura 3.67a) sia degli orari (Figura 3.67b) delle ecostazioni registrano una riduzione nella probabilità di assumere tutti gli stati ad accezione di quello massimo, cioè il 10; quest'ultimo diventa lo stato più probabile per entrambe le variabili. Migliorando

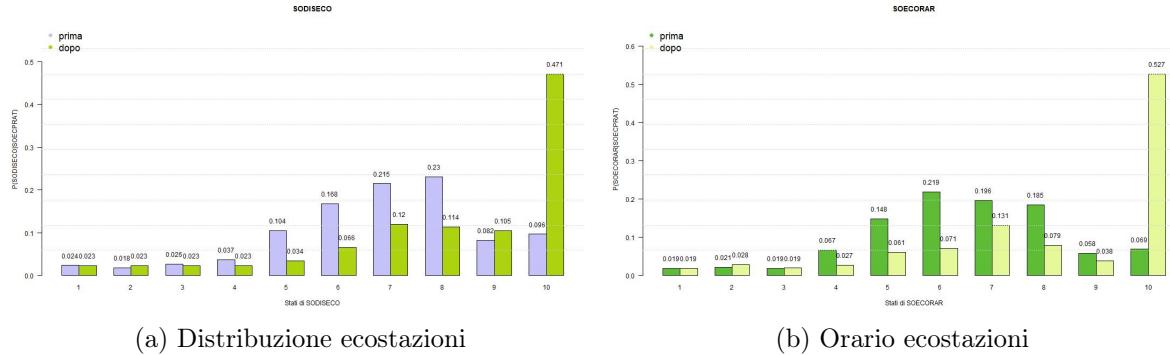


Figura 3.67: Inferenza su SOECPRAT

SOECPRAT il comune ha una buona probabilità di beneficiare di un aumento della soddisfazione legate agli altri fattori delle ecostazioni.

Alcune parole sono già state spese per segnalare la presenza di un legame tra le variabili delle ecostazioni e quelle dei container. Le variazioni riportate nei parametri dei nodi seguenti confermano le probabili sinergie che si possono ricercare in termini di soddisfazione tra container ed ecostazioni. Tutte e tre le variabili SODISCON (Figura 3.68),

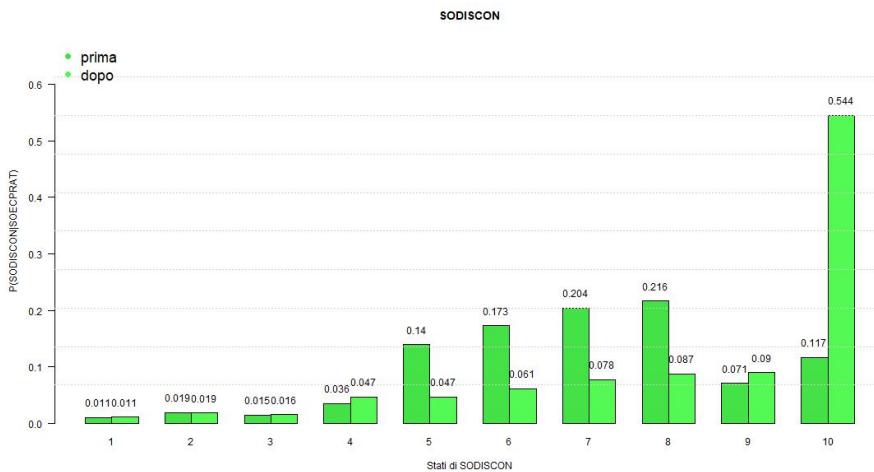


Figura 3.68: Inferenza su SOECPRAT - Distribuzione container

SOCOORAR (Figura 3.69a) e SOCOPRAT (Figura 3.69b) subiscono le stesse alterazioni nel momento in cui s'introduce l'evidenza $E \rightarrow SOECPRAT = 10$ nel MB: gli stati che indicano alti livelli di soddisfazione, 9 e 10, aumentano significativamente la propria probabilità e raggiungono insieme oltre il 50% della distribuzione.

I risultati ottenuti suggeriscono che i cittadini pienamente soddisfatti della praticità delle ecostazioni sono, in oltre il 50% dei casi anche molto soddisfatti di tutti gli aspetti legati ai container. Strategie mirate a migliorare l'esperienza dei cittadini nelle ecostazioni possono avere dei benefici anche su quest'altro elemento del servizio di raccolta rifiuti.

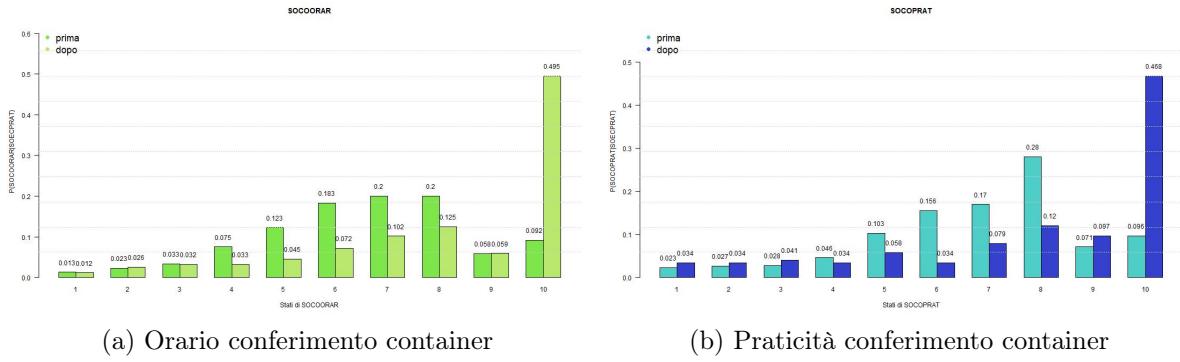


Figura 3.69: Inferenza su SOECPRAT

SOFRESER = 10 Le prime informazioni utili si traggono dalla distribuzione della variabile PROOPER (Figura 3.70). Sapendo che la soddisfazione sulla frequenza della pulizia delle strade è elevata, aumenta leggermente la probabilità che i cittadini giudichino positivamente il lavoro svolto degli operatori ecologici. Rispetto ad altri casi analizza-

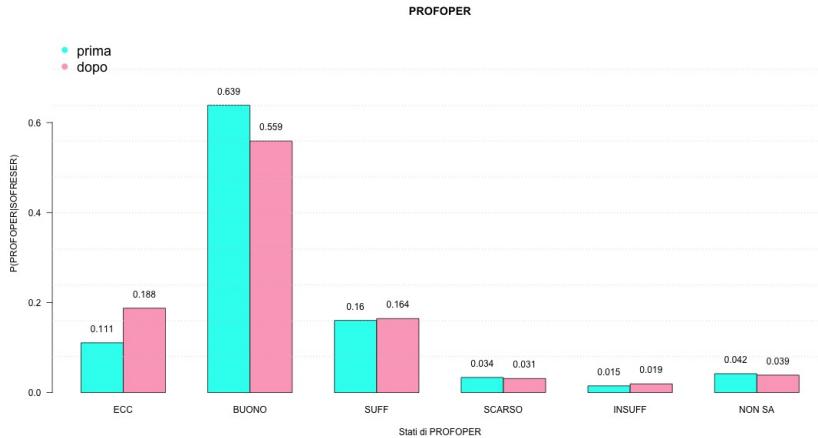


Figura 3.70: Inferenza su SOFRESER - Professionalità operatori

ti in precedenza, la variazioni assoluta è alquanto esigua, prima era $P(PROOPER = ECCELLENTE) = 0.111$ mentre ora $P(PROOPER = ECCELLENTE|SOFRESER = 10) = 0.188$. Per questo motivo, nonostante il comune possa migliorare la valutazione attribuita agli operatori attraverso SOFRESER, si possono individuare strategie più efficaci.

Ad esempio, l'influenza esercitata dalla massimizzazione di SOFRESER su SOPULSTR può costituire la base per strategie molto più efficaci rispetto alla precedente. La Figura 3.71 mostra che l'evidenza introdotta nel modello modifica radicalmente la distribuzione di probabilità della soddisfazione della pulizia delle strade. C'è una probabilità dell'83.5% che un cittadino soddisfatto completamente della frequenza del servizio, dichiari di esserlo anche della pulizia delle strade; informazioni preziosa che conferma ciò che accade nella realtà: all'aumentare della frequenza del servizio, aumenta anche la pulizia delle strade.

L'ultimo nodo appartenente allo stesso aspetto è la disposizione dei cestini (Figura 3.72a). Con l'introduzione dell'evidenza nel *MB* si assiste ad un aumento della probabilità di tutti gli stati da 7 a 10 e la corrispondente diminuzione degli stati più bassi. A differenza di quanto osservato nella maggior parte dei casi, il grado di soddisfazione più probabile è l'8 e non il massimo, cioè 10.

Legato all'aspetto della pulizia, il nodo genitore SOPULISO subisce delle variazioni

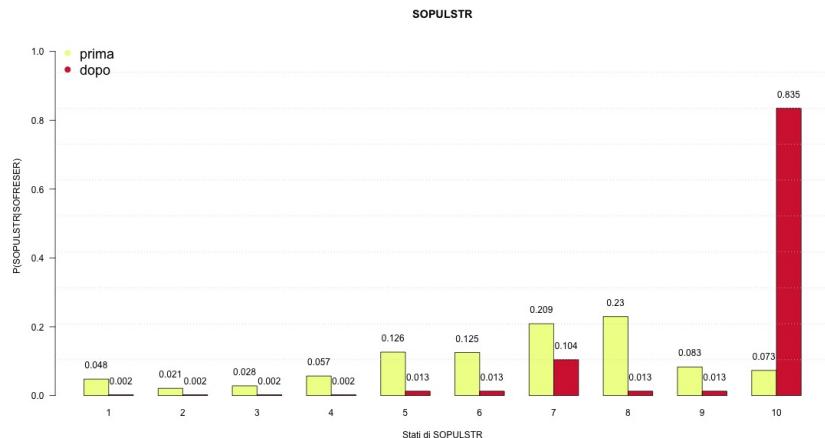


Figura 3.71: Inferenza su SOFRESER - Pulizia delle strade
maggior parte dei casi, il grado di soddisfazione più probabile è l'8 e non il massimo, cioè 10.

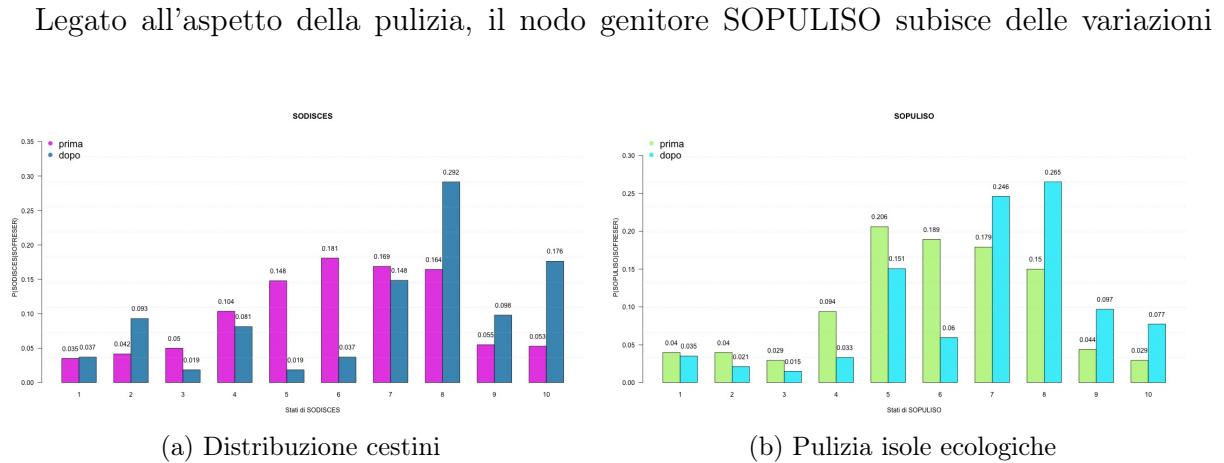


Figura 3.72: Inferenza su SOFRESER

rilevanti nella sua distribuzione di probabilità marginale (Figura 3.72b) tali che gli stati da 7 a 10 aumentano la propria probabilità a discapito degli stati indicanti gradi di soddisfazione minori, da 1 a 6.

La relazione tra SOFRESER e SOCOPRAT non è riportata in ragione dei dubbi sulla sua applicabilità in un contesto reale.

SOPULSTR = 10 Le uniche variazioni significative sono relative ai nodi SOFRESER e DECCENST, gli unici genitori di SOPULSTR. Per quanto riguarda la frequenza del servizio di pulizia (Figura 3.73a), è stato riscontrato che un'evidenza forte tipo $E \rightarrow SOPULSTR = 10$ ha un impatto estremamente positivo: la probabilità che un intervistato sia completamente soddisfatto di SOFRESER passa da $P(SOFRESER = 10) = 0.09$ a $P(SOFRESER = 10|SOPULSTR = 10) = 0.883$. L'informazione contenuta in questa relazione suggerisce che manovre comunali con effetti positivi su SOPULSTR sono probabilmente efficaci anche per aumentare il livello di soddisfazione dei cittadini sulla frequenza del servizio, perché nelle loro menti probabilmente un maggior livello di pulizia delle strade è legato significativamente al numero di volte con cui gli operatori intervengono per spazzarle.

La distribuzione di DECCENST presenta un cambiamento inaspettato: al contrario delle aspettative, sapere che un individuo è pienamente soddisfatto della pulizia delle

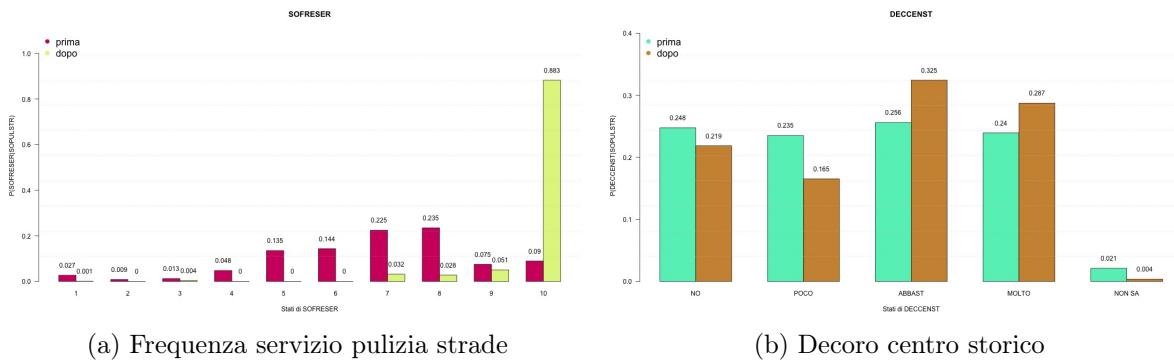


Figura 3.73: Inferenza su SOPULSTR

strade, non migliora la probabilità il suo punto di vista sulla riduzione del decoro del centro storico a causa della raccolta porta a porta. La Figura 3.73b mostra che dopo l'introduzione dell'evidenza la probabilità delle risposte positive ('No' e 'Poco') diminuisce mentre quelle negative diventano probabilmente più frequenti. Attraverso maggiori approfondimenti il comune potrebbe chiarire questo comportamento insolito.

SODISCES = 10 Con l'introduzione delle nuove informazioni nel modello, la distribuzione dell'importanza corrispondente alla disposizione dei cestini non subisce cambiamenti degni di nota. La relazione con la soddisfazione sull'orario delle ecostazioni appartiene all'insieme di relazioni da cui sembra difficile giungere a conclusioni applicabili in un contesto reale.

L'unica variabile da analizzare rimane il genitore SOFRESER (Figura 3.73a): Le pro-

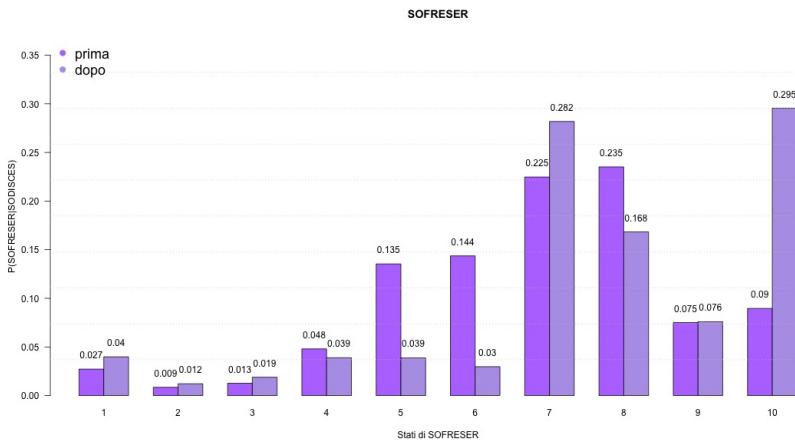


Figura 3.74: Inferenza su SODISCES - Frequenza pulizia strade

babilità associate agli stati 7, 9 e 10 aumentano dopo la propagazione delle informazioni contenute nell'evidenza, segnale che un buon livello di soddisfazione sulla distribuzione dei cestini comporta un incremento nella probabilità che anche il cittadino sia soddisfatto anche della frequenza con cui sono spazzate le strade.

3.5.2 Strategie aziendali e analisi di scenario

Nel momento in cui i *manager* devono compiere delle scelte, sono chiamati a valutare i costi ed i benefici connessi a ciascuna alternativa disponibile. Generalmente, ogni strategia ha un impatto su più di una variabile aziendale, motivo per cui è riduttivo eseguire inferenza limitandosi ai soli *Markov blanket*. Le **analisi di scenario** superano questo limite consentendo di manipolare, attraverso le evidenze, più elementi contemporaneamente e di registrare le conseguenti variazioni nella rete. Le probabili ripercussioni di ciascuna alternativa sono quantificate dai cambiamenti nei parametri della rete indotti dalle nuove informazioni inserite.

Nota tecnica: la complessità computazionale per l'elaborazione delle tabelle di probabilità condizionata sull'intera rete, è troppo elevata; per questo motivo è stata introdotta un'ipotesi semplificativa con lo scopo di ridurre le dimensioni della rete e rendere possibile lo svolgimento dell'analisi di scenario. L'ipotesi è la seguente: i 5 nodi sull'importanza globale degli aspetti, riassumono le singole importanze degli aspetti in cui sono suddivise. Questo comporta l'esclusione dalla rete dei 16 nodi sull'importanza dei fattori e l'eliminazione dei corrispondenti archi. Al termine di quest'operazione la rete, rappresentata nella Figura 3.75 è composta da 34 nodi e 66 archi direzionati.

Di seguito si propone l'analisi di alcuni scenari possibili; per ognuno di essi si evidenziano solo le variazioni più significative. Si ricorda che il comune non è in grado di manipolare alcune variabile, motivo per cui sono scartate a priori tutte le strategie che operano su queste variabili. Ad esempio: il comune non è in grado di agire sulla classe d'età o sul numero di componenti di un intervistato.

Primo scenario I costi connessi al servizio di raccolta rifiuti stanno crescendo, così il comune è costretto o ad aumentare il prelievo fiscale attraverso le tasse sui rifiuti o a diminuire il livello del servizio. L'amministrazione comunale è propensa nell'adottare la prima delle due opzioni. Per questo motivo, nel questionario è stato chiesto agli intervistati di esprimere il proprio pensiero riguardo ad un'eventuale riduzione del servizio di raccolta del secco e dell'umido con lo scopo di non dover chiedere maggiori tasse ai cittadini. Dai risultati⁸ è emerso che non tutti sarebbero favorevoli a diminuire il livello del servizio offerto.

La prima strategia da valutare richiedere che il comune interroghi i cittadini e ricerchi con loro una soluzione più accomodante in modo tale che si dichiarino tutti disposti ad accettare un passaggio in meno del servizio per il secco e l'umido; per esempio, un compromesso potrebbe essere quello di eliminare il passaggio considerato meno utile da parte dei residenti nel comune. Ipotizzando di adottare questa soluzione, s'introduce nella rete le seguenti evidenze: $RIDUSECC = 1$ e $RIDUUMID = 1$.

Propagandosi all'interno della rete, le nuove informazioni non producono scostamenti significativi nei parametri dalla rete. A titolo esemplificativo si osservi la Figura 3.76 in cui sono rappresentate le distribuzioni di probabilità marginale di due variabili SISPASTR e SODISCES, appartenenti ai *MB* dei nodi in cui è inserita l'evidenza. In termini assoluti, le variazioni sono troppo esigue per sostenere l'efficacia positiva di questa strategia.

A conferma di quanto appena detto, si osservino i *MB* dei nodi RIDUSECC (Figura 3.77a) e RIDUUMID (Figura 3.77b), entrambi di dimensioni ridotte e collegate ad un numero esiguo di nodi legati alla soddisfazione.

⁸Si veda la il paragrafo 3.2.7

Quanto emerso induce a pensare che queste variabili non abbiano una gran influenza sulla soddisfazione dei cittadini; pertanto il comune dovrebbe accantonare, almeno per il momento, questa strategia.

Secondo scenario I dati raccolti sull'importanza complessiva attribuita a ciascuno dei cinque aspetti⁹ del servizio hanno determinato che la "raccolta porta a porta" è ritenuto il più importante tra tutti. Per questo motivo, il comune potrebbe attuare una strategia per migliorare la qualità dei fattori connessi a quest'aspetto. L'aspettativa è di ottenere un miglioramento della soddisfazione globale vista l'alta importanza attribuita dai cittadini alla "raccolta porta a porta". La strategia deve essere finalizzata a migliorare tutti e quattro i fattori:

- (a) Giorni della settimana: sotto l'aspetto sia quantitativo, aumentano o diminuendo il numero di giorni, sia qualitativo, trovando le giornate più accomodanti per i cittadini;
- (b) Frequenza del servizio: prevenire, evitare o almeno ridurre il numero di volte che, per qualsiasi ragione, non si riesce ad effettuare la raccolta nei giorni previsti;
- (c) Comportamento operatori: velocità ed efficienza in fase di raccolta;
- (d) Resistenza sacchetti: aumentare la resistenza riducendo la probabilità di rompersi creando disagio ai cittadini. Ciò può essere realizzato ad esempio inspessendo i sacchetti, cambiando il materiale con cui sono realizzati, trovando nuovi fornitori.

Intraprendere una strategia come questa equivale ad inserire nella rete Bayesiana un'evidenza del tipo:

$$E \rightarrow SOGIOSET = 10, SOREGSER = 10, SOCOMOPE = 10, SORESSAC = 10$$

Le distribuzioni di probabilità dei nodi che non sono raggiunti dalle nuove informazioni rimangono invariate. I nodi in cui si propagano l'evidenze subiscono delle variazioni più o meno significative. Per cominciare si segnala un leggero miglioramento della percezione del decoro del centro storico (Figura 3.78a): migliorare la "raccolta porta a porta" aumenta le probabilità che i cittadini pensino che il centro storico non perda il proprio decoro a causa della raccolta rifiuti.

Per quanto riguarda la frequenza del servizio di spazzamento delle strade (Figura 3.78b) si rilevano variazioni anomale, nel senso che c'è miglioramento della probabilità che la soddisfazione sia massima tuttavia compensato dall'aumento della possibilità di valutazioni basse, comprese tra 1 e 3. Sono necessari maggiori approfondimenti per capire la natura di queste alterazioni nella distribuzione di probabilità.

Migliorando il servizio di "raccolta porta a porta" il comune può beneficiare di un probabile miglioramento della soddisfazione legata alla distribuzione dei cestini (figura 3.79a) e delle isole ecologiche (figura 3.79b). Le tabelle di probabilità indicano che i parametri degli stati più alti, tra 8 e 10, migliorano a discapito di quelli più bassi che subiscono una diminuzione. Questo è sicuramente un segnale positivo considerando che i nodi analizzati descrivono variabili appartenenti ad aspetti diversi da quello della "raccolta porta a porta": migliorando anche solo una parte del servizio è possibile ottenere un aumento della soddisfazione globale.

Oltre ad avere una buona influenza sulla distribuzione, l'evidenza introdotta migliora le distribuzioni di probabilità marginale degli altri due fattori relativi alle isole ecologiche, la pulizia (Figura 3.80b) e la capienza dei cassonetti (Figura 3.80a). Le variazioni sono simili a quelle descritte in precedenza: diminuiscono le probabilità d'insoddisfazione a favore di valutazioni comprese tra 8 e 10 quindi decisamente positive. Questi risultati confermano la forte relazione esistente tra l'aspetto della "raccolta porta a porta" e delle "isole ecologiche" individuata analizzando i *Markov blanket*.

Anche i nodi contenenti le importanze globali subiscono dei cambiamenti, due di esse in particolare. L'importanza della raccolta porta a porta dei rifiuti (figura 3.81a) ha un comportamento inaspettato: portando la soddisfazione del cittadino su quest'aspetto al massimo, la probabilità che egli gli attribuisca la massima importanza diminuisce. La ragione di questo comportamento si può ipotizzare sia dovuta ad un fattore psicologico per cui quando si è soddisfatti a pieno del servizio, si tende ad attribuirgli una minore importanza rispetto a prima; queste tematiche psicologiche esulano da quest'analisi pertanto ci si limita a ipotesi di questo tipo.

L'ultimo nodo contiene i dati sull'importanza globale assegnata alle isole ecologiche (Figura 3.81b). In questo caso, la propagazione dell'evidenza aumenta la probabilità che i cittadini attribuiscano a quest'aspetto del servizio un'importanza elevata, compresa tra 4 e 5.

Concludendo, attuando una strategia mirata ad aumentare la soddisfazione legata ai fattori caratterizzanti la raccolta porta a porta, il comune riuscirebbe ad aumentare il grado di soddisfazione globale. Scelte di questo tipo possono produrre sinergie utili per chi le mette in pratica poiché hanno un'ottima influenza anche sulle isole ecologiche.

Terzo scenario Servendosi nuovamente dei dati raccolti sull'importanza globale dei cinque aspetti della raccolta¹⁰, si propone l'analisi di una strategia mirata a migliorare la soddisfazione dei "container" poiché sono l'elemento ai quali gli intervistati hanno dato minor importanza.

Il piano operativo elaborato dal comune deve essere in grado di impattare positivamente sui seguenti fattori:

- (1) Distribuzione container: immaginando che sia difficile spostare le aree già esistenti, sarebbe utile investire e creare di nuove posizionandole in punti strategici per i cittadini; per determinare in quale quartiere posizionare i nuovi container si può ricercare dove risiedono gli intervistati più insoddisfatti di SODISCON;
- (2) Orario di conferimento: stabilire un arco di tempo più adeguato alle reali esigenze dei cittadini;
- (3) Praticità di conferimento: dopo aver compiuto maggiori approfondimenti per stabilire quali siano i problemi riscontrati durante il conferimento, si possono risolvere introducendo nuove procedure, nuovi container, aumentando il numero di operatori che possono agevolare il conferimento, ecc.

Implementare una strategia che abbia l'obiettivo di massimizzare questi fattori equivale ad introdurre le seguenti informazioni nella rete:

$$E \rightarrow SODISCON = 10, SOCOORAR = 10, SOCOPRAT = 10$$

⁹I dati sono riassunti nella tabella 3.10

¹⁰I dati sono riassunti nella tabella 3.10

Le tabelle di probabilità condizionata ottenute come risultato della propagazione dell'evidenza, contengono le variazioni subite dai parametri nel caso in cui si sappia che un individuo è pienamente soddisfatto dei "container". In prima battuta, osservando la Figura 3.82 si nota che i cittadini più soddisfatti risiedono, con la stessa probabilità, nel secondo, terzo e quarto quartiere; al contrario, i più insoddisfatti abitano nel primo, quinto o sesto quartiere.

Indagando sui motivi del malcontento ed agendo operativamente per migliorare il servizio offerto, il comune beneficierebbe di un incremento della soddisfazione globale.

Dai risultati ottenuti è emerso che anche le iniziative che agiscono sui "container" influiscono sulle "isole ecologiche". Tutti e tre i fattori legati a quest'ultimo aspetto subiscono variazioni positive significative, a cominciare dalla distribuzione sul territorio comunale (Figura 3.83).

Si può notare un netto miglioramento della probabilità legata ad alti valori della soddisfazione da 7 a 10. I parametri dei nodi legati alla capienza dei cassonetti (Figura 3.84a) e alla pulizia (Figura 3.84b) registrano variazioni simili. Sapere che un individuo è pienamente soddisfatto dell'organizzazione dei "container" aumenta le probabilità che lo sia anche delle "isole ecologiche". Attuando una strategia incentrata sul miglioramento dei container, il comune aumenta le probabilità che i cittadini siamo più soddisfatti anche di questo secondo aspetto.

Ad eccezione di qualche tabella di probabilità condizionata isolata, non sono avvenuti sufficienti cambiamenti nei nodi relativi agli altri elementi del servizio di raccolta rifiuti tali da giustificare una relazione d'influenza con i "container". In modo analogo, neanche le variabili sulle importanze globali hanno subito variazioni, segno che agendo sui fattori dei "container" non si possono manipolare le importanze attribuite dai cittadini ai diversi aspetti del servizio.

Il parametri del nodo SOECORAR registrano dei miglioramenti (Figura 3.85): propagandosi nella rete, l'evidenza induce un aumento della probabilità degli stati compresi tra 8 e 10 a discapito di quella corrispondente ai valori di soddisfazione più bassi. Questa relazione è spiegata dall'analisi dei *MB* eseguita in precedenza nel nella sezione 3.5.1.

Alla luce dei risultati ottenuti al termine di quest'analisi di scenario, si può concludere che una strategia incentrata sui "container", aspetto meno importante del servizio, con buona probabilità si tradurrà in un aumento della soddisfazione di altri elementi del servizio. Ciò si tradurrà in un probabile miglioramento del livello di soddisfazione globale dei cittadini.

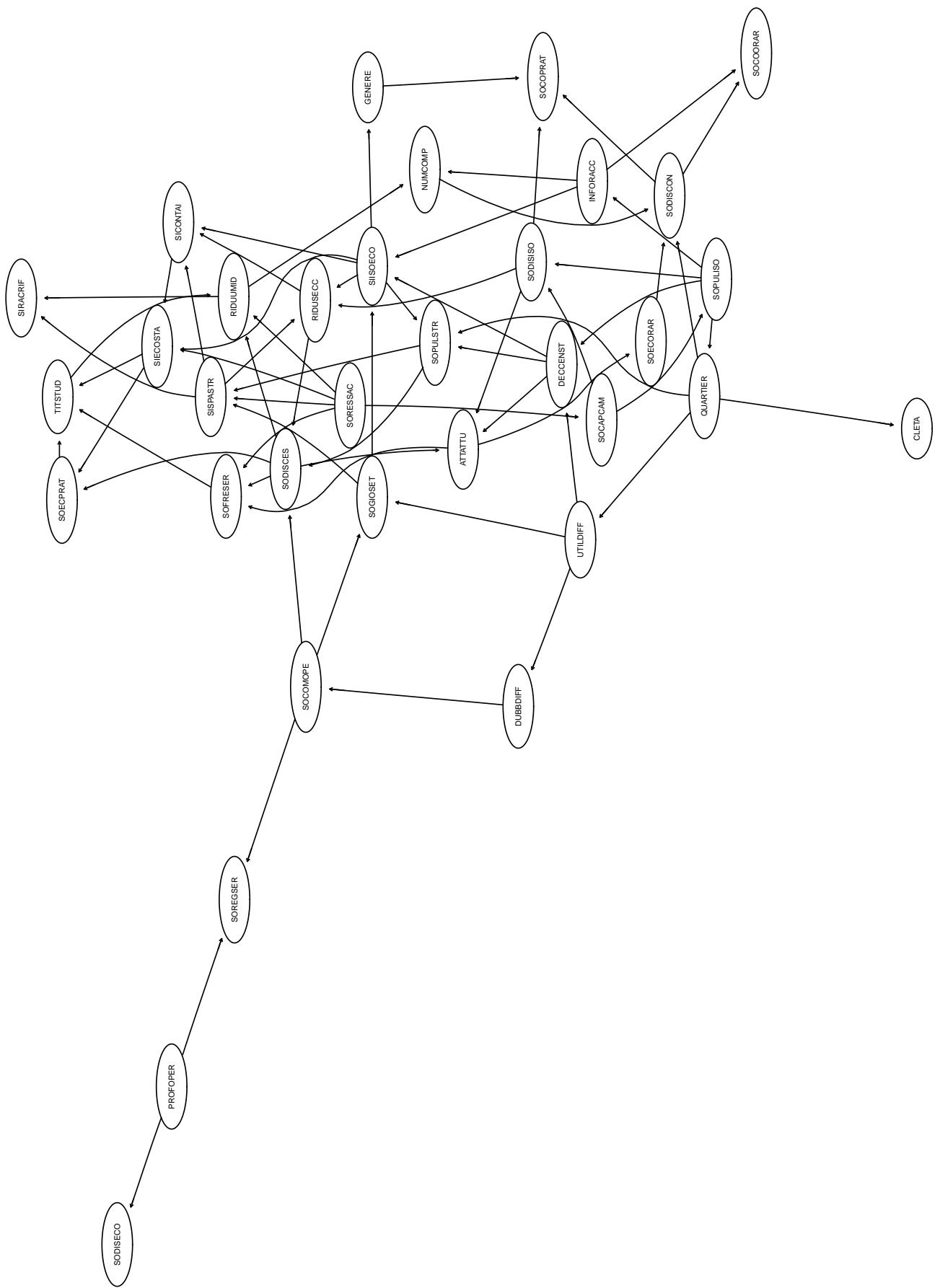


Figura 3.75: Rete Bayesiana ridotta

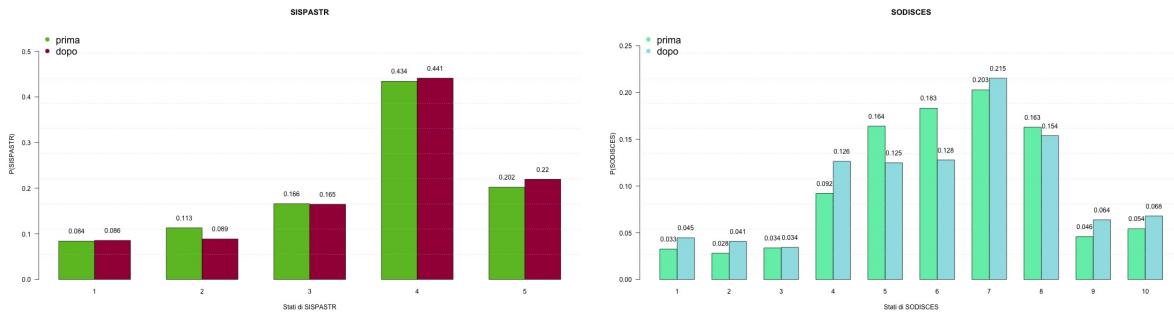


Figura 3.76: Scenario 1

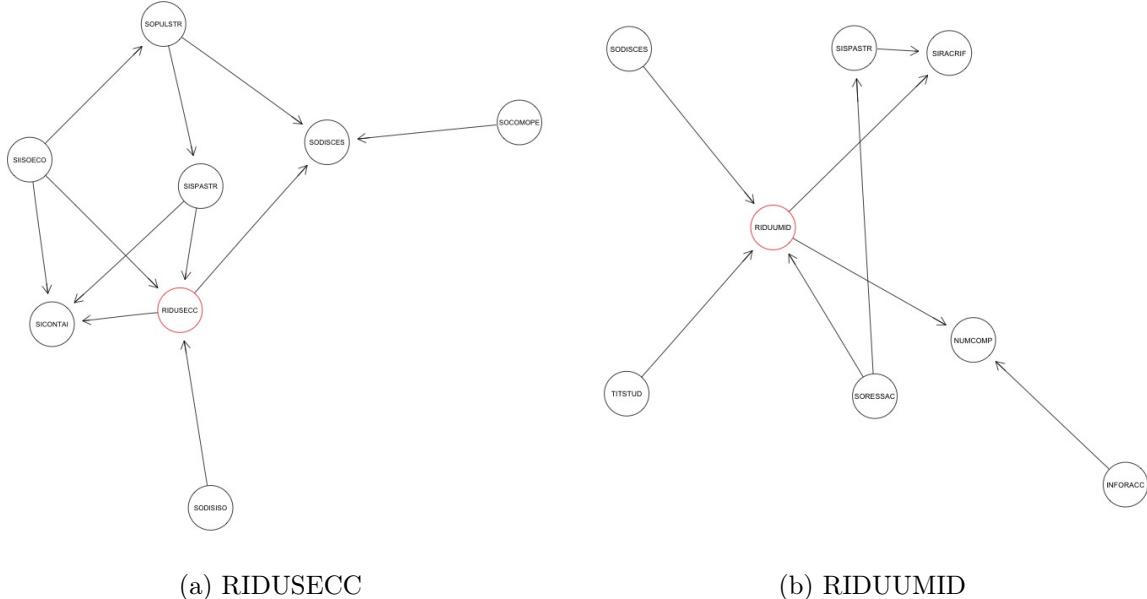


Figura 3.77: Scenario 1

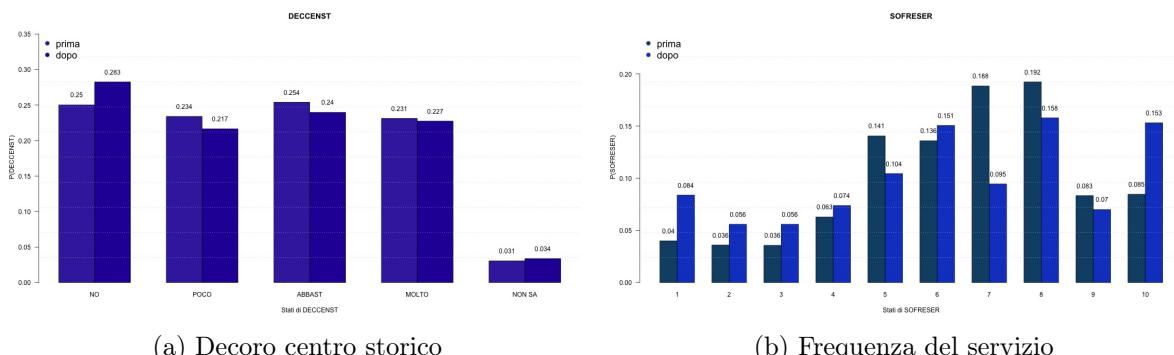


Figura 3.78: Scenario 2

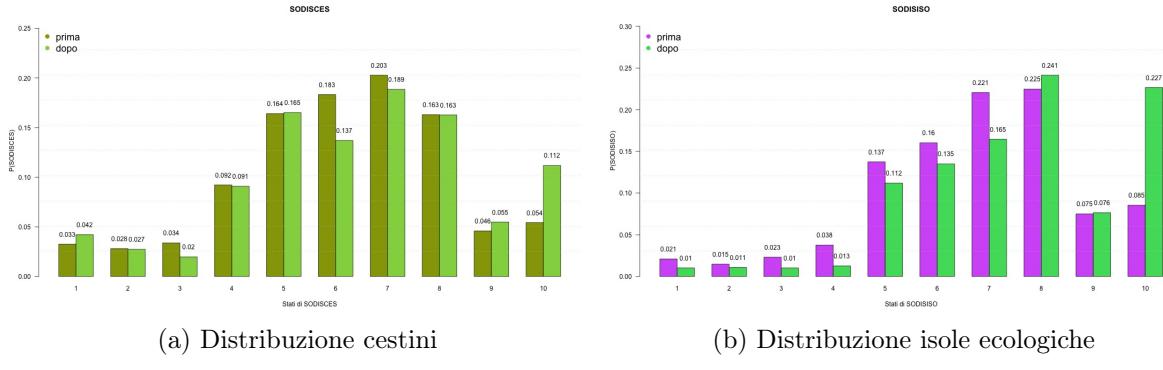


Figura 3.79: Scenario 2

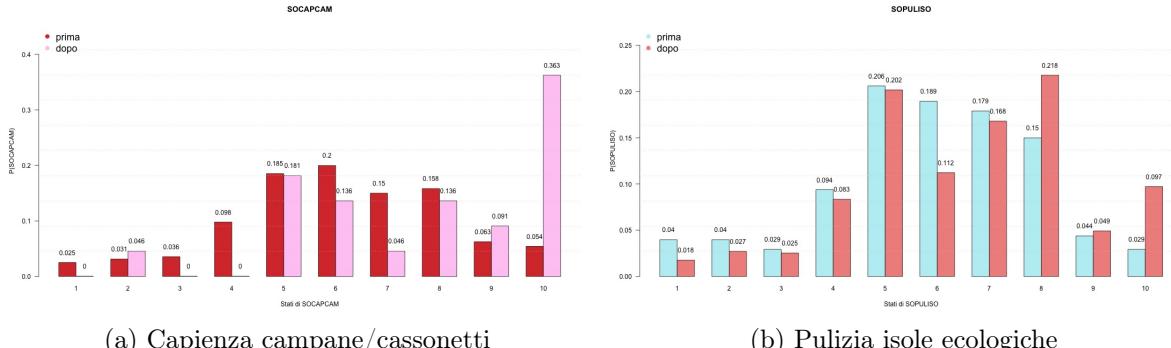


Figura 3.80: Scenario 2

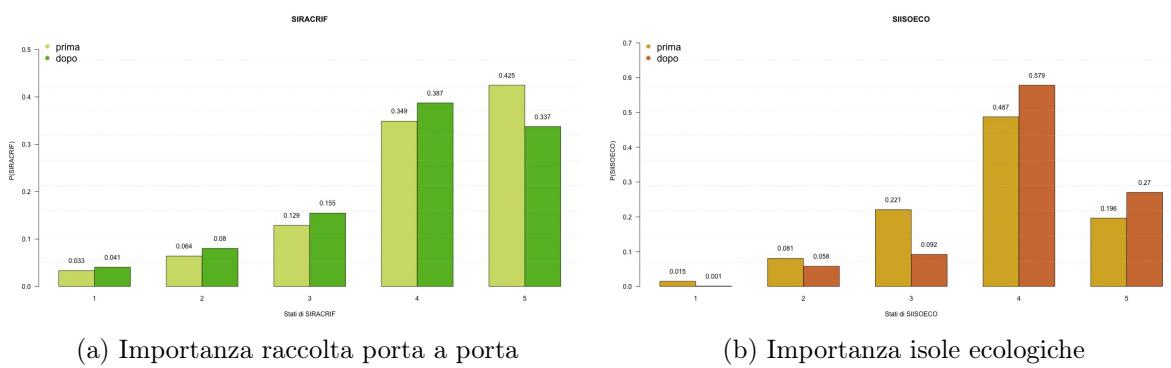


Figura 3.81: Scenario 2

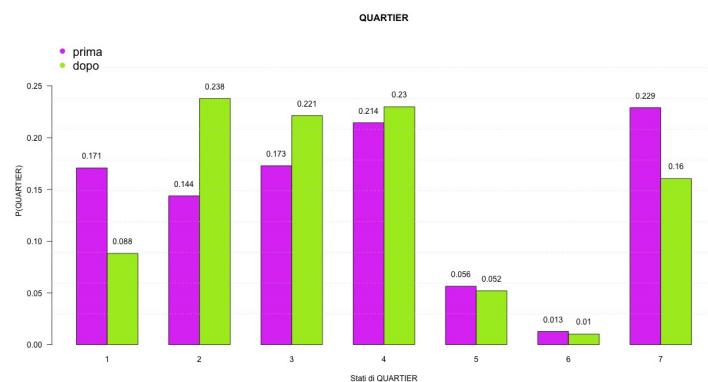


Figura 3.82: Scenario 3 - Quartiere

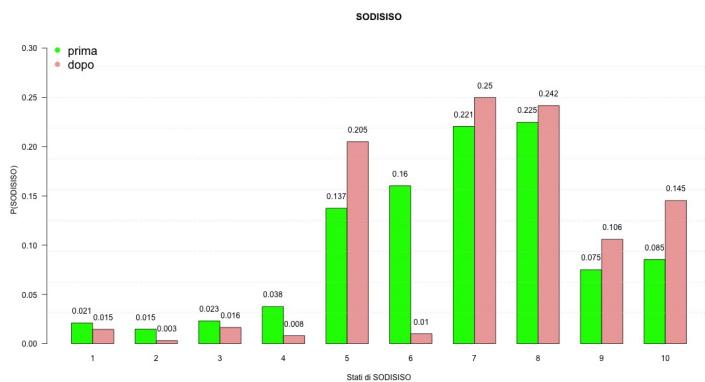


Figura 3.83: Scenario 3 - Distribuzione isole ecologiche

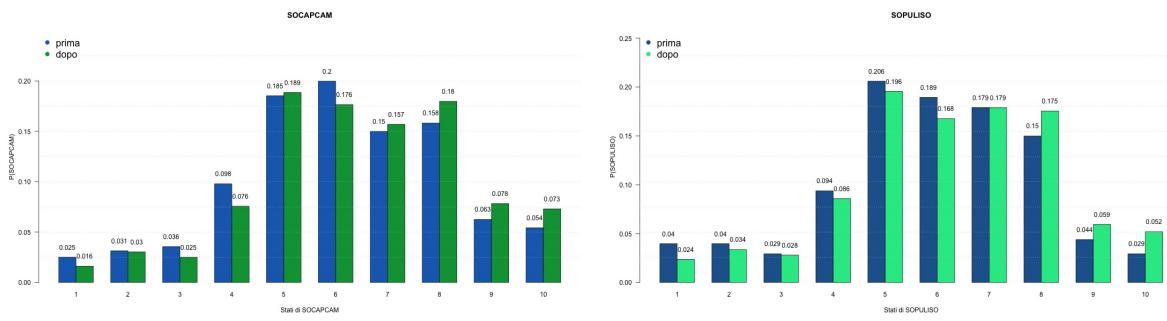


Figura 3.84: Scenario 3

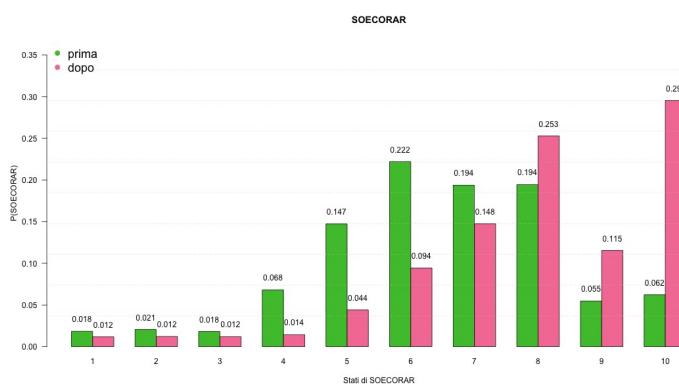


Figura 3.85: Scenario 3 - Orario conferimento ecostazioni

Conclusioni

Le informazioni ottenute dall'analisi del fenomeno oggetto di studio sono il frutto di un processo sequenziale di analisi di dati raccolti da un soggetto esterno. Attraverso dei passaggi intermedi, il *database* iniziale è stato trasformato in un insieme di dati idonei ad essere impiegati nelle elaborazioni statistiche. Per inquadrare il fenomeno oggetto di studio sono state descritte nel dettaglio le variabili contenute nella banca dati. Il passaggio successivo è stato l'apprendimento della struttura e dei parametri di una rete Bayesiana utilizzando degli appositi algoritmi di apprendimento. Poiché l'obiettivo dell'analisi è generare informazioni utili da impiegare come supporto all'interno del processo decisionale, come ultimo *step* è stato eseguito il processo d'inferenza suddiviso in due parti: (a) Analisi dei *driver* della soddisfazione di ciascun fattore; (b) Simulazione dell'impatto delle diverse strategie attraverso l'analisi di scenario. Al termine della prima fase d'inferenza, il comune è venuto a conoscenza degli elementi da modificare per massimizzare, come una certa probabilità, la soddisfazione di ciascuna fattore. Sono emerse un gran numero di relazioni significative specialmente tra le variabili riguardanti le soddisfazioni, segno che nella mente dei cittadini tutti i fattori sono più o meno connessi tra di loro. Alla medesima conclusione si è giunti anche attraverso l'analisi delle correlazioni, se pur con un minor grado di dettaglio nella descrizione delle relazioni. Di volta in volta sono stati segnalati i *driver* della soddisfazione di ciascuna caratteristica del servizio. Inoltre, sono emersi, caso per caso, tutti quei legami individuati dall'intelligenza artificiale che pur avendo un fondamento statistico non trovano un valido riscontro nella realtà. Queste relazioni, più volte richiamate, devono essere gestite con cautela.

Nella seconda fase sono state analizzate tre strategie alternative attraverso cui il comune è in grado di manipolare più variabili contemporaneamente.

La prima di esse non ha prodotto variazioni rilevanti nel livello di soddisfazione complessiva, situazione dovuta al basso numero di relazioni che coinvolgono i nodi in cui è inserita l'evidenza. Dal punto di vista pratico, questo segnale indica che nella realtà gli elementi in cui sono introdotte le nuove informazioni non influiscono sui *driver* della soddisfazione. Si sconsiglia al comune, per tanto, di ricercare una soluzione più accomodante sulla riduzione del passaggio dell'umido e del secco per evitare di aumentare le tasse sui rifiuti visto che c'è una bassa probabilità che questi sforzi siano ricompensati da un aumento della soddisfazione.

La seconda iniziativa comunale è finalizzata al miglioramento di tutti i fattori relativi alla "raccolta porta a porta". Il propagarsi dell'evidenza ha portato risultati interessanti indicando una maggior probabilità che i cittadini raggiungano un più alto livello di soddisfazione generale. Inoltre, è confermata la sinergia con le "isole ecologiche", nel senso che qualsiasi attività incentrata sulla "raccolta porta a porta", con buona probabilità, genererà un ritorno positivo anche sulle "isole ecologiche".

Nell'ultimo scenario si ipotizza di migliorare la soddisfazione dei fattori legati ai "container" che sono giudicati l'aspetto meno importante secondo gli intervistati. Le variazioni

nei parametri della rete suggeriscono che una strategia di questo tipo aumenta le probabilità che i cittadini raggiungano livelli di soddisfazioni più alti non solo per quanto riguardare le "isole ecologiche" ma anche sulla globalità del servizio.

Ragionando in termini assoluti, l'impatto più efficacie appartiene alla seconda strategia che risulta essere la migliore tra quelle analizzate. La rete Bayesiana si presta a simulare le conseguenze di qualsiasi tipo di scenario il comune voglia formulare.

Con le informazioni ottenute attraverso questo questionario non si è in grado di determinare nello specifico quali azioni è meglio intraprendere dal punto di vista pratico poiché non si conoscono le cause dell'insoddisfazione¹¹. Non si è in grado di decidere ad esempio se sia più efficacie migliorare la qualità o la quantità dei giorni in cui si effettua la raccolta; i cittadini potrebbero essere soddisfatti del numero di giorni ma non di come sono distribuiti nell'arco della settimana. Per conoscere questo genere di esigenze è necessario procedere con la raccolta di dati mirata sulla strategia che il comune è intenzionata ad intraprendere. I risultati emersi da quest'analisi sono utili per compiere una scrematura delle alternative disponibili, escludendo quelle non realistiche o poco efficienti e mettendo in luce quelle che probabilmente hanno il maggior impatto positivo sulla soddisfazioni clienti.

In questa tesi, i dati utilizzati sono relativi al grado di soddisfazione dei cittadini riguardo al servizio di raccolta rifiuti. Questo stesso procedimento empirico si presta ad essere utilizzato per esaminare le relazioni di dipendenza tra i fattori caratterizzanti qualsiasi variabile aziendale che il *management* ritiene utile analizzare; per esempio, si potrebbe considerare come il 'rischi di credito', il 'tasso di difetti' in un processo produttivo, ecc. Per concludere, le informazioni ottenute attraverso le rei Bayesiane si prestano ad essere impiegati come supporto alle decisioni, quindi non in modo esclusivo. Al contrario, la loro utilità sta nel confermare o meno, con una certa probabilità, eventuali credenze di settore, pregiudizi, conoscenze degli esperti, andamenti storici, ipotesi; il loro potenziale emerge, quindi, quando sono usate in maniera complementare assieme ad altre informazioni già in possesso.

Inoltre, dallo studio della struttura appresa si possono ricavare informazioni circa la complessa rete di relazioni che può sussistere tra le variabili che descrivono il fenomeno oggetto di studio. La rappresentazione grafica di questi modelli rende più semplice ed efficace la discussione e l'interpretazione dei risultati anche da parte di utenti che non hanno un *background* prevalentemente statistico; ciò rende le reti Bayesiane uno strumento di analisi innovativo per le aziende di qualunque settore.

¹¹Si aggiunge che le ipotesi alla base della costruzione della rete Bayesiana utilizzata in questa tesi non sono tali da poterla includere nella famiglia dei modelli causali.

Bibliografia

- S. Borra and A. Di Ciaccio. *Statistica, metodologie per le scienze economiche e sociali*. McGraw-Hill, second edition, 2008.
- S. Brasini, F Tassinari, and G. Tassinari. *Marketing e pubblicità - Metodi di analisi statistica*. Il Mulino, second edition, 1999.
- B. Busacca and E. Valdani. Customer-based view. *Finanza Marketing e Produzione*, 2: 95–131, 1999.
- S. Cenatiempo, G. D'Agostini, and A. Vanelli. Reti bayesiane: da modelli di conoscenza a strumenti inferenziali e decisionali. *Notiziario tecnico Telecom Italia*, (3):16–25, 2010.
- S. Chakraborty, C. Fidge, D. Lassen, L. Ma, and K. Mengersen. A bayesian network-based customer satisfaction model: a tool for management decisions in railway transport. *Decision Analytics*, 2016.
- G. F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.
- J. A. Cote and Joan L. Giese. Defining consumer satisfaction. *Academy of Marketing Science Review*, 2000(1), 2002.
- L.M. De Campos. A scoring function for learning bayesian networks based on mutual information and conditional independence tests. *Journal of machine learning research*, 7:2149–2187, 2006.
- E. De Jonge and M. Van Der Loo. An introduction to data cleaning with r. Technical report, Statistics Netherlands, 2013.
- A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society, B*(39):1–38, 1977.
- F. Faltin, R. Kennet, and F. Ruggeri. *Encyclopedia of Statistics in Quality and Reliability*, chapter Bayesian Networks. Wiley & Sons, 2007.
- P. A. Ferrari and G. Manzi. Nonlinear principal component analysis as a tool for the evaluation of customer satisfaction. *Quality Technology and Quantitative Management*, 7(2):117–133, 2010.
- C. M. Gonda and K. Khan. *Seal the hole in the bucket*. Embassy Books, first edition, jan 2010.
- L. Guatri. Valori di capitale economico e valori di mercato delle imprese: quali strumenti per attenuarne i divari? Università Luigi Bocconi di Milano., 1992.

-
- D. Hand, H. Mannila, and P. Smyth. *Principles of data mining (Adaptive computation and machine learning)*. Cambridge, MA: MIT Press, 2001.
- G. Iasevoli. *Il valode del cliente*. FrancoAngeli, 2010.
- R. S. Kenett and S. Salini. *Modern analysis of customer surveys - with applications using R*. John Wiley & Sons, Ltd, first edition, 2012.
- U. B. Kjaerulff and A. L. Madsen. *Bayesian network and Influence diagrams*. Springer, 2008.
- U.B. Kjaerulff. dhugin: a computational system for dynamic time-sliced bayesian networks. *International journal of forecasting*, (11):89–111, 1995.
- Kevin B. Korb and Ann E. Nicholson. *Bayesian Artifical Intelligence*. CRC Press, 2nd edition, 2011.
- M. L. Krieg. *A Tutorial on Bayesian Belief Networks*. DSTO Electronics and Surveillance Research Laboratory, 2001.
- D. Margaritis. *Learning Bayesian Network Model Structure from Data*. PhD thesis, Carnegie Mellon University, 2003.
- R. Nagarajan, M. Scutari, and S. Le'bre. *Bayesian Networks in R with Applications in Systems Biology*. Springer Science+Business Media New York, 2013.
- R. E. Neapolitan. *Learning bayesian networks*. Pearson Prentice Hall, 2004.
- J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
- J. Pearl. *Causal diagrams for empirical research*. Biometrika, 1995.
- J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, second edition, 2009.
- O. Pourret, P. Naim, and B. Marcot. *Bayesian Networks: A Practical Guide to Applications*. John Wiley & Sons, Ltd, 2008.
- J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–658, 1978.
- R. S. Salini, S. & Kenett. Bayesian networks of customer satisfaction survey data. *Journal of Applied Statistics*, 2009.
- M. Scutari and J.B. Denis. *Bayesian Networks With Examples in R*. CRC Press, 2015.
- H.A. Simon. *The New Science of Management Decision*. Harper & Row, 1960.
- L. E. Sucar. *Probabilistic graphical models*. Springer, 2015.
- C. Tarantola, P. Vicard, and I. Ntzoufras. Monitoring and improving greek banking services using bayesian networks: An analysis of mystery shopping data. *Expert Systems with Applications*, 39, 2012.
- R.A. Westbrook. Product/consumption-based affective responses and postpurchase processes. *Journal of Marketing Research*, 24(3):258–270, 1987.

Acronimi

AIC Akaike information criterion.

BDE Bayesian Dirichlet equivalent score.

BIC Bayesian information criterion.

BN rete Bayesiana.

BS Bayesian score.

CATI computer-assisted telephone interviewing.

CBV customer based view.

CPT conditional probability table.

DAG directed acyclic graph.

DBN dynamic Bayesian network.

DDN dynamic decision network.

EM expectation maximization.

HC Hill climbing.

IC Inductive causation.

ID Influence Diagram.

K2 K2 score.

kNN k nearest neighbours.

MB Markov blanket.

MCN multiply connected network.

MDL minimum description length.

ML maximum likelihood.

MLE maximum likelihood estimation.

NA Not available.

OOBN object oriented Bayesian network.

TPC tabella di probabilità condizionata.

Simboli

Ω spazio campionario.

ω insieme degli eventi elementari.

A insieme degli archi di una rete Bayesiana.

A, B, C, ... Eventi.

D database.

E evidenza.

G grafo.

M modello.

N insieme dei nodi di una rete Bayesiana.

P distribuzione di probabilità.

Pa(X_i) insieme dei genitori del nodo X_i .

V insieme delle dipendenze dirette.

X insieme delle variabili.

X, Y, Z, ... variabili casuali.

Elenco delle figure

1.1	Le cinque fasi del processo decisionale	2
2.1	Assiomi di probabilità	10
(a)	$P(A)$	10
(b)	$P(A \cup B)$	10
(c)	$P(A \cup B)$	10
2.2	Relazione diretta	12
2.3	Casi di indipendenza condizionale	15
(a)	Connessioni seriali	15
(b)	Connessioni divergenti	15
(c)	Connessioni convergenti	15
2.4	Tipologie di ragionamento	17
(a)	Diagnostico	17
(b)	Predittivo	17
(c)	Inter causale	17
(d)	Combinato	17
2.5	Struttura polialbero	20
2.6	Multiply-connected network	22
2.7	Tipologie di nodi nelle reti decisionali	31
2.8	Esempio di DBN	33
2.9	Esempio di OOBN	34
(a)	Espansione di M	34
(b)	Compressione di M	34
3.1	Genere	40
3.2	Titolo di studi	41
3.3	Attività lavorativa	41
3.4	Composizione nucleo familiare	42
3.5	Utilità raccolta	43
3.6	Informazioni sul servizio	43
(a)	Dubbi raccolta	43
(b)	Informazioni raccolta	43
3.7	Ufficio destinatario	44
3.8	Intervento operatori	45
(a)	Velocità intervento	45
(b)	Risultati intervento	45
3.9	Professionalità operatori	45
3.10	Importanza globale degli aspetti	47
(a)	Raccolta porta a porta	47

(b) Isole ecologiche	47
(c) Container	47
(d) Ecostazioni	47
(e) Pulizia strade	47
3.11 Conoscenza agevolazione	49
3.12 Propensione alla riduzione della frequenza del servizio	49
(a) Riduzione secco	49
(b) Riduzione umido	49
3.13 Decenza centro storico	50
3.14 Informazioni sui residenti nel centro storico	50
(a) Residenti nel centro storico	50
(b) Risposte residenti	50
3.15 Proposte per preservare il decoro del centro storico	51
(a) Spostamento orario	51
(b) Raccolta al mercato	51
3.16 Correlazioni	52
(a) Raccolta bisettimanale	52
(b) Isole ecologiche	52
(c) Container	52
(d) Ecostazioni	52
(e) Spazzamento e pulizia strade	52
3.17 Rete Bayesiana	55
3.18 Correlazioni confermate dalla BN	58
(a) Raccolta bisettimanale	58
(b) Isole ecologiche	58
(c) Container	58
(d) Ecostazioni	58
(e) Spazzamento e pulizia strade	58
3.19 Markov blanket	60
(a) Giorni della settimana	60
(b) Regolarità del servizio	60
3.20 Markov blanket	61
(a) Comportamento operatori	61
(b) Resistenza sacchetti umido	61
3.21 Markov blanket - Distribuzione isole ecologiche	61
3.22 Markov blanket	62
(a) Capienza cassonetti/campane	62
(b) Pulizia isole ecologiche	62
3.23 Markov blanket - Distribuzione container	63
3.24 Markov blanket	64
(a) Orario conferimento container	64
(b) Praticità conferimento container	64
3.25 Markov blanket - Orario conferimento ecostazioni	64
3.26 Markov blanket	65
(a) Distribuzione ecostazioni	65
(b) Praticità conferimento ecostazioni	65
3.27 Markov blanket - Frequenza servizio pulizia strade	66
3.28 Analisi Markov blanket	67

(a) Pulizia strade	67
(b) Distribuzione cestini	67
3.29 Inferenza su SOGIOSET - Importanza giorni raccolta	68
3.30 Inferenza su SOGIOSET - Regolarità del servizio	69
3.31 Inferenza su SOGIOSET - Riduzione passaggio umido	69
3.32 Inferenza su SOGIOSET	70
(a) Orario conferimento container	70
(b) Orario conferimento ecostazioni	70
3.33 Inferenza su SOREGSER - Giorni della settimana	70
3.34 Inferenza su SOREGSER	71
(a) Praticità container	71
(b) Comportamento operatori	71
3.35 Inferenza su SOCOMOPE - Professionalità operatori	71
3.36 Inferenza su SOCOMOPE - Regolarità del servizio	72
3.37 Inferenza su SOCOMOPE	72
(a) Praticità conferimento container	72
(b) Orario conferimento ecostazioni	72
3.38 Inferenza su SODISISO	73
(a) Distribuzione container	73
(b) Distribuzione ecostazioni	73
3.39 Inferenza su SODISISO	74
(a) Capienza cassonetti/campane	74
(b) Pulizia isole ecologiche	74
3.40 Inferenza su SOCAPCAM - Importanza capienza cassonetti	74
3.41 Inferenza su SOCAPCAM - Distribuzione isole ecologiche	74
3.42 Inferenza su SOCAPCAM - Praticità conferimento nei container	75
3.43 Inferenza su SOCAPCAM - Pulizia isole ecologiche	75
3.44 Inferenza su SOPULISO - Comportamento operatori	76
3.45 Inferenza su SOPULISO	76
(a) Capienza cassonetti/campane	76
(b) Disposizione isole ecologiche	76
3.46 Inferenza su SOPULISO - Frequenza spazzamento strade	77
3.47 Inferenza su SODISCON - Orario conferimento container	78
3.48 Inferenza su SODISCON	78
(a) Distribuzione isole ecologiche	78
(b) Distribuzione ecostazioni	78
3.49 Inferenza su SODISCON	78
(a) Praticità conferimento ecostazioni	78
(b) Orario conferimento ecostazioni	78
3.50 Inferenza su SODISCON - Decoro centro storico	79
3.51 Inferenza su SODISCON	79
(a) Giorni della settimana	79
(b) Resistenza sacchetti	79
3.52 Inferenza su SOCOORAR	80
(a) Distribuzione container	80
(b) Praticità di conferimento nei container	80
3.53 Inferenza su SOCOORAR - Distribuzione container - Importanza	80
3.54 Inferenza su SOCOORAR - Titolo di studio	81

3.55 Inferenza su SOCOORAR	81
(a) Giorni della settimana	81
(b) Orario conferimento ecostazioni	81
3.56 Inferenza su SOCOPRAT - Numero di componenti	82
3.57 Inferenza su SOCOPRAT - Orario conferimento container	82
3.58 Inferenza su SOCOPRAT	83
(a) Capienza campane/cassonetti	83
(b) Praticità conferimento ecostazioni	83
3.59 Inferenza su SOCOPRAT - Comportamento operatori	83
3.60 Inferenza su SODISECO	84
(a) Distribuzione isole ecologiche	84
(b) Distribuzione container	84
3.61 Inferenza su SODISECO	84
(a) Orario ecostazioni	84
(b) Praticità ecostazioni	84
3.62 Inferenza su SOECORAR - Attività lavorativa	85
3.63 Inferenza su SOECORAR - Orario di conferimento container	85
3.64 Inferenza su SOECORAR	86
(a) Giorni della settimana	86
(b) Regolarità del servizio	86
3.65 Inferenza su SOECORAR	86
(a) Distribuzione ecostazioni	86
(b) Praticità conferimento ecostazioni	86
3.66 Inferenza su SOECPRAT - Quartiere di residenza	86
3.67 Inferenza su SOECPRAT	87
(a) Distribuzione ecostazioni	87
(b) Orario ecostazioni	87
3.68 Inferenza su SOECPRAT - Distribuzione container	87
3.69 Inferenza su SOECPRAT	88
(a) Orario conferimento container	88
(b) Praticità conferimento container	88
3.70 Inferenza su SOFRESER - Professionalità operatori	88
3.71 Inferenza su SOFRESER - Pulizia delle strade	89
3.72 Inferenza su SOFRESER	89
(a) Distribuzione cestini	89
(b) Pulizia isole ecologiche	89
3.73 Inferenza su SOPULSTR	90
(a) Frequenza servizio pulizia strade	90
(b) Decoro centro storico	90
3.74 Inferenza su SODISCES - Frequenza pulizia strade	90
3.75 Rete Bayesiana ridotta	95
3.76 Scenario 1	96
(a) Importanza spazzamento strade	96
(b) Disposizione cestini	96
3.77 Scenario 1	96
(a) RIDUSECC	96
(b) RIDUUMID	96
3.78 Scenario 2	96

(a) Decoro centro storico	96
(b) Frequenza del servizio	96
3.79 Scenario 2	97
(a) Distribuzione cestini	97
(b) Distribuzione isole ecologiche	97
3.80 Scenario 2	97
(a) Capienza campane/cassonetti	97
(b) Pulizia isole ecologiche	97
3.81 Scenario 2	97
(a) Importanza raccolta porta a porta	97
(b) Importanza isole ecologiche	97
3.82 Scenario 3 - Quartiere	97
3.83 Scenario 3 - Distribuzione isole ecologiche	98
3.84 Scenario 3	98
(a) Capienza cassonetti	98
(b) Pulizia isole ecologiche	98
3.85 Scenario 3 - Orario conferimento ecostazioni	98

Elenco delle tabelle

3.1	Classi d'età	40
3.2	Frequenza distribuzione per classi d'età	40
3.3	Quartiere	42
3.4	Frequenza distribuzione per quartiere	42
3.5	Valutazione fattori	46
3.6	Altri servizi di raccolta	48
3.7	Distribuzione di probabilità condizionata	56
8	Significato delle variabili	113

Appendice

Appendice A: Variabili - Significato

- VARIABILE -	- SIGNIFICATO -
UTILDIFF	Utilità percepita del servizio di raccolta rifiuti.
DUBBDIFF	Dubbi sul servizio di raccolta rifiuti.
INFORACC	Sufficienza delle informazioni fornite dal comune sul servizio.
PROFOPER	Valutazione professionalità operatori.
IMGIOSET	Importanza numero giorni della settimana.
IMREGSER	Importanza regolarità del servizio.
IMCOMOPE	Importanza comportamento operatori in fase di raccolta.
IMRESSAC	Importanza resistenza sacchetti dell'umido.
IMDISISO	Importanza distribuzione isole ecologiche sul territorio comunale.
IMCAPCAM	Importanza capienza campane/cassonetti.
IMPULISO	Importanza pulizia isole ecologiche.
IMDISCON	Importanza disponibilità container sul territorio comunale.
IMCOORAR	Importanza comodità orario per il conferimento.
IMCOPRAT	Importanza praticità di conferimento.
IMDISECO	Importanza distribuzione ecostazioni sul territorio comunale.
IMECORAR	Importanza comodità orario per il conferimento.
IMECPRAT	Importanza praticità di conferimento.
IMFRESER	Importanza frequenza del servizio di spazzamento.
IMPULSTR	Importanza pulizia delle strade.
IMDISCES	Importanza distribuzione dei cestini portarifiuti sul territorio comunale.
SOGIOSET	Soddisfazione numero giorni della settimana.
SOREFSER	Soddisfazione regolarità del servizio.
SOCOMOPE	Soddisfazione comportamento operatori in fase di raccolta.
SORESSAC	Soddisfazione resistenza sacchetti dell'umido.
SODISISO	Soddisfazione distribuzione isole ecologiche sul territorio comunale.
SOCAPCAM	Soddisfazione capienza campane/cassonetti.
SOPULISO	Soddisfazione pulizia isole ecologiche.
SODISCON	Soddisfazione disponibilità container sul territorio comunale.
SOCOORAR	Soddisfazione comodità orario per il conferimento.
SOCOPRAT	Soddisfazione praticità di conferimento.
SODISECO	Soddisfazione distribuzione ecostazioni sul territorio comunale.

SOECORAR	Soddisfazione comodità orario per il conferimento.
SOECPRAT	Soddisfazione praticità di conferimento.
SOFRESER	Soddisfazione frequenza del servizio di spazzamento.
SOPULSTR	Soddisfazione pulizia delle strade.
SODISCES	Soddisfazione distribuzione dei cestini portarifiuti sul territorio comunale.
SIRACRIF	Importanza raccolta rifiuti.
SIISOECO	Importanza isole ecologiche.
SICONTAI	Importanza container.
SIECOSTA	Importanza ecostazioni.
SISPASTR	Importanza spazzamento e pulizia strade.
RIDUSECC	Disponibilità alla riduzione del passaggio del secco.
RIDUUMID	Disponibilità alla riduzione del passaggio dell'umido.
DECcenST	Decoro del centro storico.
TITSTUD	Titolo di studi.
NUMCOMP	Numero di componenti del nucleo familiare.
ATTATTU	Attività lavorativa.
GENERE	Genere.
QUARTIER	Quartieri di residenza.
CLETA	Classe d'età.

Tabella 8: Significato delle variabili

Appendice B: Questionario

QUESTIONARIO

DI RILEVAZIONE DELLA QUALITÀ PERCEPITA DAI CITTADINI PER IL SERVIZIO DI RACCOLTA RIFIUTI E PULIZIA DEL COMUNE DI XXX

L'Amministrazione Comunale di XXX, Direzione Lavori Pubblici, Servizio Ambiente, Ufficio Ecologia e verde pubblico, ha consolidato il proprio servizio di raccolta intervenendo, a partire dal giugno 2001, attraverso l'attivazione di nuove modalità di conferimento e raccolta dei rifiuti solidi urbani. I segnali di una maggiore responsabilizzazione della popolazione sono già evidenti (XXX ha dato il suo contributo con il 30% di materiali riciclati nel 2000), ma suscettibili di essere ulteriormente migliorati.

A tre anni dall'attivazione delle nuove modalità di conferimento, l'Amministrazione comunale di XXX si è posta l'obiettivo di verificare il livello di soddisfazione dei cittadini relativamente ai servizi di raccolta dei rifiuti attraverso la realizzazione di un'indagine ad hoc.

Buona sera, sono _____ della Società YYY di ZZZ. Sono stata incaricata dal Comune di XXX a svolgere un'indagine sul servizio di raccolta rifiuti urbani del Comune di XXX. Avrei bisogno di intervistare colui che, all'interno della sua famiglia, si occupa regolarmente della gestione domestica dei rifiuti.

DATI ANAGRAFICI

➤ SESSO	<input type="checkbox"/> MASCHIO	<input type="checkbox"/> FEMMINA
➤ FASCIA DI ETÀ	<input type="checkbox"/> < 25 ANNI <input type="checkbox"/> 36-45 <input type="checkbox"/> >55 ANNI	<input type="checkbox"/> 26-35 ANNI <input type="checkbox"/> 46-55 ANNI
➤ CONDIZIONE LAVORATIVA	<input type="checkbox"/> LAVORATORE DIPENDENTE <input type="checkbox"/> DISOCCUPATO <input type="checkbox"/> STUDENTE	<input type="checkbox"/> LAVORATORE AUTONOMO/IMPRENDITORE <input type="checkbox"/> PENSIONATO <input type="checkbox"/> CASALINGA
➤ COMPOSIZIONE PER NUCLEO FAMIGLIARE	<input type="checkbox"/> SINGLE <input type="checkbox"/> FAMIGLIA CON 3 COMPONENTI <input type="checkbox"/> FAMIGLIA CON PIÙ DI 4 COMPONENTI	<input type="checkbox"/> FAMIGLIA CON 2 COMPONENTI <input type="checkbox"/> FAMIGLIA CON 4 COMPONENTI
➤ TITOLO DI STUDIO CONSEGUITO	<input type="checkbox"/> SCUOLA DELL'OBBLIGO <input type="checkbox"/> MEDIE SUPERIORI	<input type="checkbox"/> LAUREA

INFORMAZIONE SUL SERVIZIO

1. LEI RITIENE CHE LA RACCOLTA DIFFERENZIATA SIA UTILE PER (FORNIRE UNA SOLA RISPOSTA)	
➤ DIMINUIRE LA QUANTITÀ DI RIFIUTI DA AVVIARE A SMALTIMENTO	<input type="checkbox"/>
➤ RIDURRE GLI SPRECHI DI RISORSE E MATERIE PRIME	<input type="checkbox"/>
➤ FAR RISPARMIARE IL COMUNE	<input type="checkbox"/>
➤ MIGLIORARE IL DECORO DELLA CITTÀ	<input type="checkbox"/>
➤ NON NE VEDO L'UTILITÀ	<input type="checkbox"/>
➤ NON SO – NON RISPONDE	<input type="checkbox"/>

2. HA ANCORA DUBBI SU COME EFFETTUARE LA RACCOLTA DIFFERENZIATA

- | | |
|-------------------------|--------------------------|
| ➤ SI, MOLTI | <input type="checkbox"/> |
| ➤ SI, ALCUNI | <input type="checkbox"/> |
| ➤ NO, NESSUNO | <input type="checkbox"/> |
| ➤ NON SO – NON RISPONDE | <input type="checkbox"/> |

3. RITIENE CHE LE INFORMAZIONI FORNITE SUL SERVIZIO DI RACCOLTA RIFIUTI SIANO SUFFICIENTI

- | | |
|-------------------------|--------------------------|
| ➤ SI | <input type="checkbox"/> |
| ➤ NO | <input type="checkbox"/> |
| ➤ NON SO – NON RISPONDE | <input type="checkbox"/> |

RUOLO OPERATORI**4. HA MAI PROPOSTO SUGGERIMENTI O EFFETTUATO SEGNALAZIONI O RECLAMI RELATIVAMENTE AL SERVIZIO DI RACCOLTA RIFIUTI**

- | | |
|------|--------------------------|
| ➤ SI | <input type="checkbox"/> |
| ➤ NO | <input type="checkbox"/> |

[PROSEGUIRE SOLO IN CASO DI RISPOSTA AFFERMATIVA ALLA DOMANDA 4]

5. SE SI, A QUALE UFFICIO SI È RIVOLTO

- | | |
|---|--------------------------|
| ➤ UFFICIO AMBIENTE DEL COMUNE | <input type="checkbox"/> |
| ➤ UFFICI DELLA DITTA CHE EFFETTUÀ LA RACCOLTA | <input type="checkbox"/> |
| ➤ ALTRI UFFICI DEL COMUNE | <input type="checkbox"/> |

6. IN SEGUITO AL RECLAMO/SEGNALAZIONE HA RICEVUTO RISPOSTA?

- | | |
|------|--------------------------|
| ➤ SI | <input type="checkbox"/> |
| ➤ NO | <input type="checkbox"/> |

7. LE SEGNALAZIONI O I RECLAMI HANNO VISTO UN INTERVENTO DEGLI ADDETTI DEL COMUNE O DEGLI OPERATORI ADDETTI AL SERVIZIO DI RACCOLTA RIFIUTI

- | | |
|---------------------|--------------------------|
| ➤ TEMPESTIVO | <input type="checkbox"/> |
| ➤ RITARDATO | <input type="checkbox"/> |
| ➤ NESSUN INTERVENTO | <input type="checkbox"/> |

8. I RISULTATI DELL'INTERVENTO DEGLI ADDETTI DEL COMUNE O DEGLI OPERATORI ADDETTI AL SERVIZIO DI RACCOLTA RIFIUTI SONO STATI

- | | |
|--------------|--------------------------|
| ➤ EFFICACI | <input type="checkbox"/> |
| ➤ PARZIALI | <input type="checkbox"/> |
| ➤ INEFFICACI | <input type="checkbox"/> |

9. COME VALUTA IL GRADO DI PROFESSIONALITÀ NEGLI OPERATORI ADDETTI AL SERVIZIO DI RACCOLTA RIFIUTI

- | | |
|--------------|--------------------------|
| ➤ ECCELLENTE | <input type="checkbox"/> |
|--------------|--------------------------|

➤ BUONO	<input type="checkbox"/>
➤ SUFFICIENTE	<input type="checkbox"/>
➤ SCARSO	<input type="checkbox"/>
➤ INSUFFICIENTE	<input type="checkbox"/>

VALUTAZIONE DEL SERVIZIO

Le chiedo ora di analizzare insieme una serie di “**fattori**”, che descrivono in dettaglio il servizio di igiene urbana e di assegnare il *livello di importanza* utilizzando una scala da 1 a 5 (dare 5 al fattore ritenuto più importante e assegnare un voto da 1 a 4 ai restanti fattori); e assegnare il *livello di soddisfazione*, utilizzando una scala da 1 a 10 (come a scuola) dove il 10 rappresenta l'eccellenza (per X fattore la Vostra soddisfazione è totale), il 6 la sufficienza (per X fattore non siete completamente soddisfatti), 4 (per X fattore non siete soddisfatti) ecc.

ASPETTO	FATTORI	IMPORTANZA Attribuire voti da 1 a 5	SODDISFAZIONE Attribuire voti da 1 a 10
➤ RACCOLTA BISETTIMANALE PORTA A PORTA DI SECCO E UMIDO	Numero giorni raccolta a settimana Regolarità del servizio ¹ Comportamento operatori in fase di raccolta Resistenza sacchetti raccolta umido	1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
➤ ISOLE ECOLOGICHE punti adibiti alla raccolta di carta, vetro, lattine e plastica	Distribuzione isole ecologiche sul territorio del comune Capienza campane/cassonetti ² Pulizia isole ecologiche	1 2 3 4 5 1 2 3 4 5 1 2 3 4 5	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
➤ CONTAINER adibiti alla raccolta rifiuti ingombranti /vegetali presso i quartieri	Disponibilità container sul territorio del comune Comodità orario conferimento Praticità di conferimento	1 2 3 4 5 1 2 3 4 5 1 2 3 4 5	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
➤ ECOSTAZIONI aree attrezzate appositamente nelle zone periferiche	Distribuzione ecostazioni sul territorio del comune Comodità orario conferimento Praticità di conferimento	1 2 3 4 5 1 2 3 4 5 1 2 3 4 5	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
➤ SPAZZAMENTO E PULIZIA STRADE	Frequenza del servizio Pulizia delle strade Distribuzione dei cestini portarifiuti sul territorio del comune	1 2 3 4 5 1 2 3 4 5 1 2 3 4 5	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

¹ Inteso come rispetto dei giorni di raccolta da parte degli operatori

² Spazio disponibile nella fase di conferimento dei rifiuti (pieno, vuoto)

In sintesi quindi, secondo Lei tra gli aspetti analizzati che descrivono il servizio d'igiene urbana quale è il più importante?

Dare 5 al fattore ritenuto più importante e assegnare un voto da 1 a 4 ai restanti fattori

<ul style="list-style-type: none"> ➤ RACCOLTA PORTA A PORTA DI SECCO E UMIDO ➤ ISOLE ECOLOGICHE punti adibiti alla raccolta di carta, vetro, lattine e plastica ➤ CONTAINER adibiti alla la raccolta rifiuti ingombranti /vegetali presso i quartieri ➤ ECOSTAZIONI Aree attrezzate di Magrè, piscine e zona industriale ➤ SPAZZAMENTO E PULIZIA DELLE STRADE 	1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5
--	---

CONOSCENZA E UTILIZZO DEGLI ALTRI SERVIZI DI RACCOLTA OFFERTI DAL COMUNE

QUALI SERVIZI DI RACCOLTA OFFERTI DAL COMUNE CONOSCE O UTILIZZA			
	NON CONOSCO	CONOSCO MA NON UTILIZZO	UTILIZZO
➤ DISTRIBUZIONE GRATUITA COMPOSTER DOMESTICI	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
➤ FORNITURA KIT PER MICRORACCOLTA AMIANTO	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
➤ RACCOLTA PORTA A PORTA DI VETRO/CARTA PER ATTIVITÀ COMMERCIALI	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
➤ RACCOLTA PORTA A PORTA DI CARTA PER UFFICI ZONA INDUSTRIALE	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
➤ RACCOLTA TONER, CARTUCCE ESAURITE PER STAMPANTI c/o UTENZE COMMERCIALI, UFFICI COMUNALI	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
➤ RACCOLTA DI PILE ESAUSTE PRESSO ALCUNE UTENZE COMMERCIALI	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
➤ RACCOLTA FARMACI SCADUTI PRESSO LE FARMACIE	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
➤ PULIZIA PARCHI GIOCHI	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
➤ PULIZIA CADITOIE/TOMBINI STRADALI	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
➤ CONSEGNA KIT BIDONCINO BLU PER L'AVVIO DELLA RACCOLTA SECCO-UMIDO	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

LIVELLO DI CONTRIBUZIONE

10. È A CONOSCENZA DELL'AGEVOLAZIONE DEL 20% SULLA TASSA DEI RIFIUTI NEL CASO EFFETTUI IL COMPOSTAGGIO DOMESTICO (SMALTIMENTO IN PROPRIO DELL'UMIDO)

- | | |
|-------------------------|--------------------------|
| ➤ SI | <input type="checkbox"/> |
| ➤ NO | <input type="checkbox"/> |
| ➤ NON SO – NON RISPONDE | <input type="checkbox"/> |

IN FUTURO, CONSEGUENTEMENTE AD OBBLIGHI LEGISLATIVI, PER GARANTIRE GLI ATTUALI SERVIZI POTREBBE ESSERE NECESSARIO AUMENTARE LA TASSA PER LO SMALTIMENTO DEI RIFIUTI URBANI.
PUR DI MANTENERE INVARIATO L'ATTUALE LIVELLO DI TASSAZIONE LEI SAREBBE DISPOSTO

11. AD UNA RIDUZIONE DEL PASSAGGIO DELLA RACCOLTA DEL SECCO AD UNA VOLTA A SETTIMANA

- | | |
|-------------------------|--------------------------|
| ➤ SI | <input type="checkbox"/> |
| ➤ NO | <input type="checkbox"/> |
| ➤ NON SO – NON RISPONDE | <input type="checkbox"/> |

IN FUTURO, CONSEGUENTEMENTE AD OBBLIGHI LEGISLATIVI, PER GARANTIRE GLI ATTUALI SERVIZI POTREBBE ESSERE NECESSARIO AUMENTARE LA TASSA PER LO SMALTIMENTO DEI RIFIUTI URBANI.
PUR DI MANTENERE INVARIATO L'ATTUALE LIVELLO DI TASSAZIONE LEI SAREBBE DISPOSTO

12. AD ELIMINARE IL TERZO PASSAGGIO DI RACCOLTA DELL'UMIDO PREVISTO NEL PERIODO ESTIVO

- | | |
|-------------------------|--------------------------|
| ➤ SI | <input type="checkbox"/> |
| ➤ NO | <input type="checkbox"/> |
| ➤ NON SO – NON RISPONDE | <input type="checkbox"/> |

SERVIZIO DI RACCOLTA CON MODALITÀ PORTA A PORTA NEL CENTRO STORICO

PER TUTTO IL CAMPIONE

1. RITIENE CHE LE MODALITÀ DI RACCOLTA PORTA A PORTA DI SECCO E UMIDO IMPATTINO SUL DECORO DEL CENTRO STORICO	
➤ PER NIENTE	<input type="checkbox"/>
➤ POCO	<input type="checkbox"/>
➤ ABBASTANZA	<input type="checkbox"/>
➤ MOLTO	<input type="checkbox"/>
➤ NON SO – NON RISPONDE	<input type="checkbox"/>

PER IL SOLO CAMPIONE RIFERITO AL CENTRO STORICO

NEL CASO SI RISPOSTA AFFERMATIVA (ABBASTANZA E MOLTO) PER MIGLIORARE IL DECORO DEL CENTRO

2. SAREBBE FAVOREVOLE AD UNA VARIAZIONE DELLE MODALITÀ DEL PORTA A PORTA CHE COMPORTI LO SPOSTAMENTO DELL'ORARIO DI RACCOLTA ALLA FASCIA MATTUTINA DELLE 8.00-10.00	
➤ PER NIENTE	<input type="checkbox"/>
➤ POCO	<input type="checkbox"/>
➤ ABBASTANZA	<input type="checkbox"/>
➤ MOLTO	<input type="checkbox"/>

3. TROVEREBBE UTILE CONFERIRE I RIFIUTI ANCHE NEI PUNTI DI RACCOLTA DISPONIBILI NEI GIORNI DI MERCATO	
➤ PER NIENTE	<input type="checkbox"/>
➤ POCO	<input type="checkbox"/>
➤ ABBASTANZA	<input type="checkbox"/>
➤ MOLTO	<input type="checkbox"/>

Software

L’analisi è svolta utilizzando il linguaggio di programmazione **R**¹² un ambiente software gratuito per il calcolo statistico e la creazione di grafici ed disponibile per la maggior parte delle piattaforme UNIX, Windows e MacOS. *R* fornisce un’ampia varietà di tecniche statistiche (modellizzazione lineare e non lineare, test statistici classici, analisi di serie temporali, classificazione, *clustering*, ...) e tecniche grafiche, inoltre è altamente estensibile grazie ai molti pacchetti installabili disponibili.

Il *package* **VIM**¹³ è stato impiegato nella fase di *data cleaning* per svolgere l’operazione di imputazione degli NA attraverso l’algoritmo *k nearest neighbours*.

Tra tutti i pacchetti utilizzati, il più importante è **bnlearn**¹⁴ attraverso cui è stato possibile l’apprendimento della struttura e dei parametri della BN. *bnlearn* offre un’ampia varietà di algoritmi d’apprendimento della struttura (approcci sia *constraint-based*, sia *score-based* con la possibilità di elaborare strutture discrete, continue o ibride), approcci d’apprendimento dei parametri (*maximum likelihood* per dati discreti e continui, *Bayesian score (BS)* per dati discreti) e tecniche di inferenza (convalida incrociata, *bootstrap*, *query* di probabilità condizionale e previsione). Sul piano tecnico si segnala che è anche l’unico pacchetto che mantiene una chiara separazione tra la struttura di una rete e la distribuzione di probabilità associata, che sono implementate come due diverse classi di oggetti R. Per approfondimenti su questo *package* si consulti uno dei due libri dedicati alle reti Bayesiane e alla loro elaborazione con R: (a) *Bayesian Networks - With Examples in R* (Scutari and Denis, 2015); (b) *Bayesian Networks in R with Applications in Systems Biology* (Nagarajan et al., 2013).

Per svolgere il processo d’inferenza sono state impiegate le funzioni contenute nel *package* **gRain**¹⁵, le quali implementano gli algoritmi per la propagazione della probabilità nelle reti Bayesiane.

Ovviamente, i *software* appena citati non esauriscono l’insieme dei programmi che consentono di analizzare un fenomeno attraverso le reti Bayesiane. Non solo sono disponibili altri *package* per l’ambiente R come *catnet* o *deal*, ma esistono anche altre piattaforme in grado di lavorare con le BN, tra queste ci sono *BayesiaLab* o *Hugin*. Per avere una panoramica di tutti questi *software* con le relative caratteristiche si consulti (Scutari and Denis, 2015, pp.125)

¹²<https://www.r-project.org/>

¹³<https://cran.r-project.org/web/packages/VIM>

¹⁴<http://www.bnlearn.com/>

¹⁵<https://cran.r-project.org/web/packages/gRain>

Ringraziamenti

Metaforicamente parlando, questa tesi rappresenta la linea d'arrivo del mio percorso di studi, traguardo che difficilmente sarei riuscito a raggiungere se fossi stato da solo in questa corsa. Per mia fortuna ho avuto di fianco a me molte persone che mi hanno accompagnato e quindi voglio dedicare a loro le ultime parole di questa tesi.

I primi in assoluto sono sicuramente le persone che mi hanno cresciuto e dato la possibilità di studiare, supportandomi ed aiutandomi a trovare la grinta per affrontare qualsiasi sfida; sto parlando di Michele e Nadia, i miei genitori. Parlando di famiglia, voglio e devo ringraziare di cuore anche mio fratello Matteo e mia sorella Mara per cui provo un grande affetto. Tra le persone a me vicine ci sono gli zii Malvina e Dino, i cugini Matteo e Luca, il nonno Sergio. Dedico questa tesi anche ai nonni Natalina e Mario che vorrei potessero condividere con me questo momento.

Se durante questo percorso non sono mai stato solo, è merito dei molti amici che mi sono stati vicino, per questo mi auguro di non dimenticare nessuno. Quelli che conosco da più tempo solo gli "Azzano people", quindi Giorgia, Enrica, Marco, Martina e soprattutto Veronica, che mi parlano ancora nonostante tutte le volte che gli ho "tirato pacco". Da anni ormai io, Federico M. ed Alessandra ci impegniamo a ritagliare una sera della settimana per stare assieme; per questo li ringrazio sperando che anche in futuro tutto questo possa continuare. Al contrario, Alice e Chiara (Kia) sono amiche che per colpa delle distanze vedo di rado, nonostante ciò trovano sempre il modo di farmi sentire la loro presenza ed il loro appoggio. Oltre dieci anni fa ho sviluppato un'enorme passione per la pallavolo, sport di squadra che oltre ad avermi fatto conoscere moltissime persone, mi ha cambiato più di una volta la vita. Per questo motivo ringrazio tutti i compagni con cui ho condiviso e condiviso lo spogliatoio almeno quattro volte a settimana, e tutti gli allenatori che hanno contribuito alla mia crescita non solo sportiva ma anche personale. Un grazie particolare a Federico B. con cui ho coltivato un'amicizia anche fuori dalle linee del campo da gioco. Restando nel mondo dello sport, ringrazio Lisa, Mattia, Nicola e Samuele per la loro compagnia durante i tornei estivi. *The last, but not least*, i compagni con cui ho letteralmente percorso questo cammino di studio, Maria, Quaira, Matteo e Tommaso, per essere stato mio coinquilino anche se solo per un anno.

Un doveroso ringraziamento anche alla mia relatrice, la prof.ssa Slanzi per avermi aiutato e seguito con gran disponibilità durante tutta la stesura di questa tesi, in cui ho potuto esprimere la mia passione sia per la statistica che per l'informatica. Non posso dimenticare anche il dott. Tomaselli e tutto lo *staff* di Demetra s.r.l. che oltre ad avermi permesso di svolgere da loro il periodo di tirocinio mi hanno fornito i dati per sviluppare questa tesi. Grazie ancora a tutti per aver contribuito a questo traguardo.

Marco