Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

La socialità degli ingredienti: analisi della composizione di ricette con modelli per reti

Relatrice Prof. Mariangela Guidolin
Dipartimento di Scienze Statistiche

Laureanda: Caterina Sgarbossa
Matricola N 2026886

Anno Accademico 2022/2023

Indice

Introduzione **1**

1 I dati di rete **4**
1.1 Caratteristiche e proprietà delle reti 7
1.2 Statistiche descrittive 9

2 Modelli per l'analisi di dati di rete **12**
2.1 Modelli ad effetti additivi 13
2.1.1 Scomposizione ANOVA e Social Relation Model 14
2.1.2 Social Relation Regression Model 16
2.2 Modelli ad effetti latenti 16
2.2.1 Latent Eigenmodel 17
2.2.2 Additive and Multiplicative Effects Model . . 18

3 I dati **21**
3.1 Le ricette 21
3.2 Gli ingredienti 25

4 Le reti di ingredienti **28**
4.1 I composti organici 29

4.2 L'informazione mutua puntuale 33

5 La modellazione della rete **38**

5.1 Social Relation Regression Model 40

5.2 Additive and Multiplicative Effects Model 43

6 Conclusioni **48**

Bibliografia **51**

A Il codice **55**

Introduzione

Questo lavoro nasce da un articolo pubblicato sul *Scientific Reports* da Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow e Albert-László Barabási [1], all'interno del quale gli autori cercano i criteri secondo i quali diverse culture alimentari considerano buono un piatto. L'approccio scelto dagli autori è di descrivere le cucine attraverso una rete sociale in cui l'oggetto di studio sono gli ingredienti presenti nei piatti, con lo scopo di individuare delle regole per abbinare in maniera efficace gli ingredienti all'interno di una ricetta. Nello specifico, l'obiettivo finale è quello di testare l'ipotesi dell'abbinamento di sapori [2], secondo la quale l'abbinare ingredienti con sapori simili porta ad un risultato finale migliore.
La strada intrapresa passa per una descrizione accurata delle caratteristiche delle cucine considerate, ottenuta attraverso la network analysis. Per fare ciò, è stata usata una rete che registra, per ogni coppia di ingredienti, il numero di composti organici in comune tra i 1 021 considerati [3]. Tale scelta è motivata dal fatto che, tra tutti i fattori che concorrono nel determinare il sapore di un ingrediente, siano proprio i composti organici a svolgere un ruolo centrale [4]. I risultati presentati sono concentrati su due aree: Nord America ed

Asia orientale. Lo studio degli ingredienti e dei loro abbinamenti porta gli autori a concludere che le differenze regionali sono troppo profonde per permettere di definire delle regole di gusto globali ma permettono di individuare le preferenze all'interno delle diverse cucine.

Al di là delle conclusioni tratte, questo articolo si trova alla base di molti dei lavori che, negli anni successivi, sono andati a comporre quella che viene definita gastronomia computazionale [5]. Si tratta di un lavoro che è stato in grado di fornire un approccio e una struttura per numerosi lavori: dalla caratterizzazione delle cucine tramite clustering [6] alla creazione di nuove ricette tramite sistemi di raccomandazione [7].

Si decide quindi di effettuare un'analisi con lo stesso obiettivo dell'articolo originario, sviluppandola attraverso l'uso di indici e modelli statistici, con l'intento di verificare se tali strumenti sono in grado di cogliere degli ulteriori aspetti.

Per contestualizzare la motivazione che ha portato all'utilizzo dell'analisi di rete, all'interno del primo capitolo vengono fatti dei cenni storici all'origine delle reti sociali, ripercorrendo il loro sviluppo fino alle più recenti applicazioni, a cui seguono delle nozioni di base della teoria delle reti e gli indici descrittivi di maggiore interesse in presenza di reti indirette e pesate, contesto in cui si colloca la casistica considerata.

Vengono poi presentati, nel secondo capitolo, dei modelli additivi quali il Social Relation Model e il Social Relation Regression Model, ed i modelli ad effetti latenti Latent Eigenmodel e Additive and Multiplicative Effects Model, scelti per la loro capacità di cogliere le complesse dipendenze riscontrabili in presenza di variabili diadiche [8].

In seguito alle analisi preliminari e le elaborazioni necessarie per rendere i dati in grado di fornire le informazioni richieste - le quali vengono svolte nel terzo capitolo - si decide di incentrare l'analisi sulle ricette provenienti dalle aree sulle quali ha incentrato la presentazione dei propri risultati lo studio con cui si effettua il confronto, per mantenere una certa coerenza.
Le analisi esplorative, descritte nel quarto capitolo, suggeriscono che la relazione tra il numero di composti organici che gli ingredienti condividono e il criterio per il loro utilizzo all'interno delle ricette sia diversa tra le varie cucine regionali. Nel capitolo successivo, lo studio viene affinato attraverso la presentazione e la valutazione dei modelli SRRM e AME, i quali presentano un soddisfacente adattamento ai dati e forniscono delle stime utili a rispondere alla domanda di ricerca. Tali stime identificano dei comportamenti simili tra le due cucine, contraddicendo quanto evidenziato nella prima fase dell'analisi. Tuttavia, viene registrata anche una certa dose di incertezza attraverso l'individuazione di fattori latenti e pertanto di effetti non spiegabili attraverso i dati disponibili. Questo porta a concludere con alcune proposte sul tipo di dati che potrebbero integrare le informazioni disponibili e, grazie all'adeguatezza dimostrata dagli strumenti statistici utilizzati, rendere possibile il raggiungimento di un maggiore grado di certezza.

Capitolo 1

I dati di rete

Qualora l'interesse di una ricerca sia quello di indagare le relazioni tra gli elementi di un insieme - o risulti necessario tenerne conto all'interno di un'analisi più ampia - l'analisi delle reti sociali risulta essere un potente strumento.
Si tratta di una disciplina la cui impostazione teorica si trova già negli scritti dei sociologi Georg Simmel [9] e Émile Durkheim [10], alla fine del diciannovesimo secolo. Sono loro, infatti, i primi ad utilizzare il termine "rete sociale" per denotare il complesso insieme di relazioni che legano i membri di un sistema sociale.
La necessità di uno studio sistematico del concetto di rete nasce successivamente, nel secondo dopoguerra [11]. A causa dei profondi cambiamenti legati a quel periodo storico la società interesse di studio risulta caratterizzata non solo da grandi dimensioni, ma anche da particolare eterogeneità e mobilità, in una maniera completamente nuova.
Quella che ne risulta è una nuova visione: si inizia ad intendere l'essere umano come un essere sociale in grado di influire sul comportamento degli altri ed allo stesso tempo di venirne influenzato. Rappresentazione che grazie allo sviluppo della teoria dei grafi diventa

non più una mera idea, ma un concetto analitico.
Storicamente, l'esperimento condotto dallo psicologo statunitense Stanley Milgram [12] è la più famosa applicazione di questo concetto. Il suo obiettivo è quello di verificare l'ipotesi che il mondo è "piccolo", cioè che ogni persona al mondo possa essere raggiunta, tramite una rete di amicizie, con relativamente pochi passaggi. Milgram affida qualche centinaio di lettere a delle persone selezionate in maniera causale, residenti ad Omaha, Nebraska, e chiede loro di spedirle ad una predefinita persona a Boston, Massachusetts. La particolarità consiste nel fatto che l'indirizzo di questa persona è ignoto ai partecipanti all'esperimento. Quindi, i soggetti si ingegnano per spedire la lettera alla persona che secondo loro ha la maggior probabilità di conoscere il destinatario finale e chiedono di ripetere l'operazione fino al raggiungimento dell'obiettivo. Contrariamente ad ogni aspettativa - non sarebbe strano supporre che per compiere un'impresa simile servano centinaia di passaggi - il numero medio di intermediari registrati è sei, da cui prende il nome quella che è conosciuta come la teoria dei sei gradi di separazione [12].
Questo esperimento può portare a pensare che la struttura di una rete sia abbastanza semplice e sia possibile spostarsi facilmente all'interno di essa. Tuttavia si tratta di una semplificazione ottimista nella maggior parte dei casi. Rimanendo nella teoria dei sei gradi di separazione, infatti, si può pensare di spiegarla come segue: se ogni soggetto ha cento amici, che a loro volta ne hanno cento, tramite un semplice calcolo si può concludere che ogni persona ha quasi dieci miliardi di amici di quinto grado. Pertanto, se ogni persona ha cento amici può connettersi con chiunque altro nel mondo con sei passaggi. Tuttavia questo ragionamento presenta una chiara falla, è infatti poco realistico pensare che i cento amici del primo soggetto preso in

considerazione non siano tra i cento amici gli uni degli altri. Questo permette di capire quanto sembri un paradosso quello del mondo piccolo, e di conseguenza quanto forte possa essere l'approfondito studio delle reti sociali come strumento.

A teorizzare e formalizzare l'idea di grafo, sono i matematici Paul Erdős ed il suo collaboratore, Alfréd Rényi. Il loro principale contributo è quello di dimostrare l'esistenza di un punto critico nei grafi casuali [12]. Quello che evidenzia il loro lavoro è che, appena superata la soglia della singola connessione media per elemento del sistema, i vertici del grafo tendono ad essere tutti collegati tra loro. Questo risultato è essenziale in quanto, se due attori non sono in alcun modo legati tra loro non possono influenzarsi, quindi potrebbero anche non far parte della stessa rete. Siccome a priori si assume che all'interno di un sistema non ci sia una regola, ma le connessioni siano casuali, e l'obiettivo è quello di dimostrare che in realtà una regola nascosta c'è, questo risultato risulta essere essenziale. Sarebbe impossibile lavorare su una rete senza intenderla come un sistema interconnesso, in quanto, anche continuando ad aggiungere collegamenti, certe parti del sistema non entrerebbero mai in contatto tra di loro.

Un'altra implicazione importante dell'introduzione di un rigore matematico all'interno di una disciplina astratta è la possibilità di applicarla in vari ambiti: quelle che ormai si chiamano reti sociali solo si nome trovano oggi applicazione contesto, dalla quantificazione della reputazione di un artista [13] allo studio di come una guerra possa influenzare la disponibilità di cibo a livello globale [14].

1.1 Caratteristiche e proprietà delle reti

Per rete si intende un gruppo di unità, detti *nodi*, caratterizzato da un insieme di connessioni, dette *archi*, che forniscono una misura di quale sia l'interazione tra una coppia di nodi, a cui ci si riferisce usando il termine *diade*.
Una rete viene comunemente rappresentata tramite un grafo $G = (N, A)$, dove $N = \{1, \ldots, V\}$ è l'insieme dei nodi, mentre $A \subseteq N \times N$ è l'insieme degli archi. La cardinalità dell'insieme di vertici è detta *ordine*, mentre la variabile associata alla presenza di un arco viene detta *variabile diadica*.
Uno strumento di rappresentazione delle reti è la *matrice di adiacenza* $\mathbf{Y}$. Si tratta di una matrice di dimensioni $V \times V$ la cui componente $y_{i,j}$ indica il valore della relazione considerata tra il nodo i e il nodo j dalla prospettiva del nodo i. Nello specifico

$$[\mathbf{Y}]_{i,j} = \begin{cases} y_{i,j} = 0 & \text{se non è presente un arco tra } i \text{ e } j, \\ y_{i,j} \neq 0 & \text{se } i \text{ presenta una relazione con } j. \end{cases}$$

In base alle caratteristiche di questa matrice si hanno vari tipi di rete a cui corrispondono grafi differenti.
Una prima distinzione viene fatta in base alla simmetria della matrice. Se $\mathbf{Y}$ è simmetrica, cioè $y_{i,j} = y_{j,i}$, la relazione tra i nodi è reciproca e si parla di rete *indiretta* che viene rappresentata da un grafo *non orientato*. Se invece il valore cambia in base al ruolo dei nodi in quanto mittenti o destinatari, cioè $y_{i,j} \neq y_{j,i}$, ci si trova di fronte ad una rete *diretta*, alla quale corrisponde un grafo *orientato*.

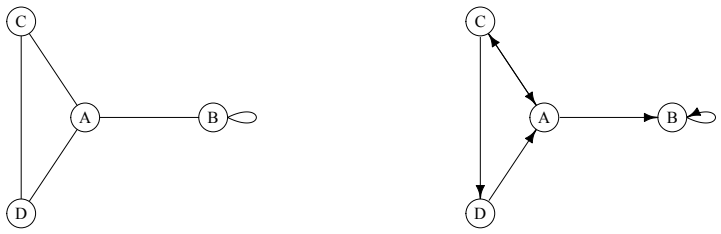

Figura 1.1: Un esempio rispettivamente di grafo non orientato e di grafo orientato

La matrice di adiacenza del grafo presentato in figura è

$$\mathbf{Y} = \begin{bmatrix} 0 & y_{A,B} & y_{A,C} & 0 \\ 0 & y_{B,B} & 0 & 0 \\ y_{C,A} & 0 & 0 & y_{C,D} \\ y_{D,A} & 0 & 0 & 0 \end{bmatrix}$$

la quale, in base alla valorizzazione degli archi, fornisce delle ulteriori distinzioni tra i vari tipi di reti esistenti.
Se il valore sulla diagonale $y_{B,B}$ è valorizzato diverso da zero la relazione è *riflessiva*, in caso contrario si avrebbe una diagonale identicamente nulla, ad indicare una relazione *antiriflessiva*.
Invece, la presenza di valori $y_{i,j} \in \{0, 1\}$, $\forall i, j \in N$, implica la presenza di una la matrice che si limita a registrare la presenza della variabile diadica o la sua assenza. In tal caso si parla di una rete *binaria*. Contrariamente, se i valori appartengono all'asse reale e forniscono un'indicazione dell'intensità della relazione, la rete è *pesata* e la matrice di adiacenza viene indicata col termine *sociomatrice*.
Nel caso di matrici sparse può risultare conveniente rappresentare la relazione tramite una *lista d'adiacenza*: una lista contenente tutte le coppie di vertici per cui esiste un arco. Nel caso di una rete pesata

la lista può essere integrata una terca colonna contenente il valore della relazione.
Nel seguito verranno presentate delle nozioni assumendo di trovarsi in presenza di una relazione pesata, indiretta e non riflessiva.

1.2 Statistiche descrittive

All'aumentare della cardinalità dei nodi N, le nozioni presentate fin'ora risultano insufficienti a descrivere in maniera soddisfacente la natura della rete. Diventa pertanto necessario l'utilizzo di indici utili a descrivere le caratteristiche dei nodi che popolano la rete e diano una visione d'insieme delle complesse relazioni a cui danno origine:

- *Grado*: numero di nodi a cui il nodo i-esimo è connesso, cioè

$$n_i = \sum_{j=1}^{V} I_{\mathbb{R}\setminus 0}\left(y_{i,j}\right).^{1} \tag{1.1}$$

  Si tratta di un indice che indica quanto il nodo si trova coinvolto nelle relazioni della rete, gli elementi caratterizzati da gradi particolarmente elevati e con una funzione di ponte all'interno della struttura sociale sono detti *hub*.

  Nel caso di reti pesate risulta interessante il calcolo di una versione pesata del grado, detta *forza* [15] del nodo i, definita come

$$s_i = \sum_{j=1}^{V} y_{i,j}, \tag{1.2}$$

  somma dei pesi degli archi che comprendono l'elemento i.

[1] Funzione indicatrice pari ad 1 se l'arco $y_{i,j}$ è valorizzato non nullo, 0 altrimenti.

- *Densità*: si tratta della frequenza relativa del numero di archi osservati sul totale degli archi possibili, cioè

$$D = \frac{1}{V(V-1)} \sum_{i,j \in N} I_{\mathbb{R}\setminus 0}\left(y_{i,j}\right), \tag{1.3}$$

  quantifica quanto è connessa la rete considerata.

- *Coefficiente di clustering*: se la presenza di una relazione tra i e j e j e k comporta una connessione anche tra i e k allora si dice che la triade gode della proprietà di transitività. La presenza di molte triadi transitive comporta la formazione di cluster, pertanto ci si riferisce a questa proprietà anche come *transitività* e viene calcolata come

$$T = \frac{\sum_{i,j,k \in N} I_{\mathbb{R}\setminus 0}\left(y_{i,j} \cdot y_{i,k} \cdot y_{j,k}\right)}{\sum_{i,j,k \in N} I_{\mathbb{R}}\left(y_{i,j} \cdot y_{i,k} \cdot y_{j,k}\right)}, \tag{1.4}$$

  la quale restituisce il rapporto tra tutte le triadi transitive e il numero totale di triadi possibili, restituendo così un valore normalizzato.

  Per le reti pesate, è stato sviluppato anche il concetto di *coefficiente di clustering pesato* [15], col quale risulta possibile calcolare l'indice per ogni nodo oltre che valore globale, dato dalla media aritmetica dei coefficienti ottenuti per i singoli elementi. In letteratura ne esistono due versioni normalizzate particolarmente diffuse ed utilizzate [16].

  La prima versione si calcola per ogni nodo i come

$$C_i^+ = \frac{1}{s_i(n_i - 1)} \sum_{j,k \in N} \frac{y_{i,j} + y_{i,k}}{2} I_{\mathbb{R}\setminus 0}\left(y_{i,j} \cdot y_{i,k} \cdot y_{j,k}\right) \tag{1.5}$$

  dove s_i è la forza del nodo i e n_i è il suo grado. Non tenendo conto del valore del terzo arco della triade ma solo della sua

esistenza, questa formulazione risulta maggiormente concentrata sul comportamento del singolo nodo i nella formazione di triadi. Diversamente, considerando il coefficiente nella forma

$$C_i^{\times} = \frac{1}{n_i(n_i-1)} \sum_{j,k \in N} (\hat{y}_{i,j}\hat{y}_{i,k}\hat{y}_{j,k})^{\frac{1}{3}} \tag{1.6}$$

dove

$$\hat{y}_{j,k} = \frac{y_{j,k}}{\max(y_{l,h})_{l,h \in N}}.$$

L'uso dei pesi dell'intera triade rende l'indice sensibile anche al peso degli archi negli intorni del nodo i e non solo alla loro presenza.

Capitolo 2

Modelli per l'analisi di dati di rete

Per la modellazione di una variabile diadica è possibile l'utilizzo di un modello lineare generalizzato stimato tramite massima verosimiglianza [17], in cui la distribuzione di probabilità da cui i dati sono generati $y_{i,j} \sim \mathbb{P}(Y \mid \theta_{i,j})$ assume ogni osservazione indipendente dalle altre dato

$$\theta_{i,j} = \beta^\top \mathbf{x}_{i,j},$$

dove $\mathbf{x}_{i,j}$ è un vettore di regressori. Ad esempio, si potrebbe avere $\mathbf{x}_{i,j} = (x_i, x_{i,j})$ dove x_i è un vettore di caratteristiche del nodo i e $x_{i,j}$ è un vettore delle caratteristiche della coppia (i, j).
Nello specifico viene costruita una funzione di densità congiunta su tutte le diadi nella forma

$$\mathbb{P}(\mathbf{Y} \mid \theta) = \prod_{i=1}^{n} \prod_{j=1}^{n} \mathbb{P}(y_{i,j} \mid \theta_{i,j}) = L(\theta \mid \mathbf{Y}). \tag{2.1}$$

La verosimiglianza così definita è valida solo sotto l'assunzione di indipendenza di $y_{i,j}$ da $y_{j,i}$ e $y_{i,k}$ condizionatamente alla covariate considerate.

Considerare $y_{i,j}$ e $y_{j,i}$ indipendenti coincide con affermare l'assenza reciprocità nelle relazioni, cosa insensata nei numerosi frangenti in cui gli eventi sono solo reciproci. Porre $y_{i,j}$ e $y_{i,k}$ indipendenti risulta ancora più difficile da giustificare in quanto equivale a supporre che la relazione tra i e j non sia influenzata in alcun modo da come i si relaziona con k e come k si lega a j. Le relazioni rappresentate dai dati di rete, infatti, si trovano contestualizzate all'interno di un sistema più ampio. Ne segue che l'esistenza di dipendenze diadiche o a livello di nodo potrebbero portare ad avere delle forme di dipendenze tra eventi anche in parti diverse della rete.
Ne segue che i modelli che ignorano potenziali dipendenze tra le osservazioni diadiche spesso riportano stime distorte e scarse capacità predittive, da cui la necessità di sviluppare modelli che ne tengano conto, come quelli proposti da Peter D. Hoff [18].

2.1 Modelli ad effetti additivi

Tra le possibili dipendenze che si possono riscontrare analizzando una sociomatrice, la più comune è quella di primo ordine - a livello di nodo - nella quale i valori della variabile diadica di una riga siano correlati tra loro. Questa dipendenza è spiegabile tramite la socialità del mittente e comporta che si avrà una certa omogeneità tra le relazioni di un singolo nodo, cioè i valori $y_{i,j}$ e $y_{i,l}$ saranno più simili tra di loro di quanto non lo siano con qualsiasi altro $y_{h,k}$ con $h \neq i$. La presenza di bassi ed alti valori non equidistribuiti tra le righe si traduce in un'eterogeneità tra le medie di riga.
Analogamente si caratterizza la popolarità del nodo in quanto destinatario. Quindi, così come l'eterogeneità dei nodi in termini della loro socialità contribuisce alla varianza tra le medie di riga della so-

ciomatrice, la popolarità contribuisce alla varianza tra le medie di colonna.
Inoltre spesso si nota come i nodi più sociali siano anche i più popolari, pertanto le medie di riga e colonna della matrice di adiacenza potrebbero essere correlate. Ovviamente, nel caso di una rete indiretta gli effetti di riga e di colonna coincidono e si parla solo di un effetto a livello di nodo.

2.1.1 Scomposizione ANOVA e Social Relation Model

Per valutare l'eterogeneità tra righe e colonne si può usare il modello basato sulla scomposizione ANOVA, il quale assume che la variabilità degli $y_{i,j}$ attorno ad una media totale μ sia ben rappresentata dagli effetti additivi di riga e di colonna [18]:

$$y_{i,j} = \mu + a_i + b_j + \varepsilon_{i,j} \tag{2.2}$$

Questo modello coglie l'eterogeneità tra gli a_i (effetto del mittente) e i b_i (effetto del destinatario), fornendo così l'eterogeneità osservata rispettivamente nelle medie delle righe e delle colonne della sociomatrice.
Il principale limite di questo modello consiste nel fatto che venga ignorato come ogni nodo sia presente sia come mittente che come destinatario. Infatti ogni nodo i ha due diversi effetti additivi: un effetto di riga a_i e un effetto di colonna b_i. Dato che ogni coppia di effetti (a_i, b_i) condivide un nodo risulta naturale aspettarsi una correlazione tra i vettori $(a_1, \ldots, a_n)$ e $(b_1, \ldots, b_n)$, contrariamente a come avviene nel modello appena presentato. Inoltre ogni diade (i, j) ha due valori associati, $y_{i,j}$ e $y_{j,i}$ la cui correlazione viene indicata col termine reciprocità. Pertanto andrebbe perlomeno

considerata la possibilità dell'esistenza di una correlazione tra $\varepsilon_{i,j}$ e $\varepsilon_{j,i}$.

Le dipendenze appena descritte, che l'ANOVA risulta incapace di cogliere, sono dette di secondo ordine o diadiche.

Un modello in grado di tener contro delle dipendenze di primo e secondo ordine è il Social Relation Model (SRM).

Esso scompone la varianza delle osservazioni di una matrice di adiacenza in termini di eterogeneità tra le medie delle righe, eterogeneità tra le medie delle colonne, correlazone tra le medie di riga e colonna e correlazione tra le diadi [17]:

$$y_{i,j} = \mu + a_i + b_j + \varepsilon_{i,j} \tag{2.3}$$

dove

$$(a_1, b_1), \ldots, (a_n, b_n) \overset{i.i.d.}{\sim} N(0, \Sigma_{ab})$$
$$\{(\varepsilon_{i,j}, \varepsilon_{j,i}) \mid i \neq j\} \overset{i.i.d.}{\sim} N(0, \Sigma_{\varepsilon})$$

con

$$\Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \qquad \text{e} \qquad \Sigma_{\varepsilon} = \sigma_{\varepsilon}^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Ne segue che la variabile che descrive la relazione risulta avere varianza

$$\text{Var}\,(y_{i,j}) = \sigma_a^2 + 2\sigma_{ab} + \sigma_b^2 + \sigma^2 \tag{2.4}$$

e covarianze

$$\text{Cov}\,(y_{i,j}, y_{i,k}) = \sigma_a^2 \qquad \text{(covarianza tra righe)}$$
$$\text{Cov}\,(y_{i,j}, y_{k,j}) = \sigma_a^2 \qquad \text{(covarianza tra colonne)}$$
$$\text{Cov}\,(y_{i,j}, y_{j,k}) = \sigma_{ab} \qquad \text{(covarianza tra righe e colonne)}$$
$$\text{Cov}\,(y_{i,j}, y_{i,k}) = 2\sigma_{ab} + \rho\sigma^2 \qquad \text{(covarianza tra righe e colonne più reciprocità)}$$

pari a zero in tutti gli altri casi.
In questa notazione gli effetti di riga e colonna sono modellati assieme per tenere conto della correlazione tra quanto un nodo è attivo in quanto mittente e destinatario. Le eterogeneità di riga e colonna sono colte rispettivamente da σ_a^2 e σ_b^2, mentre σ_{ab} descrive la reazione lineare tra i due effetti (i.e. se un nodo più sociale è anche più popolare). Le dipendenze del secondo ordine sono descritte da σ_ε^2, così come la correlazione diadica (reciprocità) è modellata dal parametro ρ.

2.1.2 Social Relation Regression Model

Spesso risulta interessante quantificare la relazione tra la variabile oggetto di studio e un'altra diadica o una variabile nodale. Per fare ciò si stima un Social Relations Regression Model (SRRM) [18], combinazione di un modello di regressione lineare con la struttura di covarianza del modello SRM [17] che diventa:

$$y_{i,j} = \beta^\top \mathbf{x}_{i,j} + a_i + b_j + \varepsilon_{i,j} \tag{2.5}$$

dove $\mathbf{x}_{i,j}$ è un vettore p-dimensionale di regressori (covariate della diade e dei nodi in quanto mittenti e destinatari) e β è il vettore dei coefficienti da stimare.

2.2 Modelli ad effetti latenti

Qualora la presenza di caratteristiche condivise da sottoinsiemi di nodi influenzi la probabilità dell'interazione tra di loro ci si trova a trattare dipendenze di terzo ordine. Una situazione simile è spesso caratterizzata da un grafo con un alto numero di triadi transitive, cioè terne di nodi in cui ogni nodo è collegato agli altri due.

L'implicazione importante quando ci si trova a fare inferenza è che, a meno che non ci sia la possibilità di quantificare tutte le variabili in grado di spiegare la presenza di queste triadi, la probabilità che i e j formino un legame non è indipendente dai legami già esistenti tra questi due nodi e un altro elemento della rete. In tal caso si parla di transitività, solitamente riscontrabile qualora ci si trovi in presenza di *omofilia*, intesa come tendenza dei nodi a creare legami con nodi che presentano caratteristiche simili.

2.2.1 Latent Eigenmodel

Le dipendenze di terzo ordine possono essere descritte da un modello a con effetti moltiplicativi, così da considerare la combinazione di effetti latenti in maniera non additiva, come accade per l'*equivalenza stocastica* [19]. Due nodi i e j sono detti essere stocasticamente equivalenti se la probabilità che i presenti una connessione con di ogni altro nodo è pari a quella di j. Questo concetto fa riferimento all'idea che ci saranno nodi all'interno della rete con pattern relazionali simili.

Per catturare questo genere di struttura risulta necessario l'utilizzo di un modello a fattori latenti, il quale assume che la relazione tra i nodi sia mediata da un piccolo numero (K) di variabili di nodo non osservate.

Tra i modelli di questa natura risulta particolarmente interessante il modello basato sugli autovalori [20], il quale viene usato per descrivere dati relazionali simmetrici, assumendo che la relazione d'interesse sia funzione del prodotto interno pesato di due vettori di variabili di nodo latenti.

Il modello descrive $y_{i,j}$ come funzione di

$$\beta^\top \mathbf{x}_{i,j} + \mathbf{u}_i^\top \Lambda \mathbf{u}_j \tag{2.6}$$

dove $\mathbf{u}_l \in \mathbb{R}^K$ con $l \in \{1, \ldots, n\}$ sono dei fattori a livello di nodo e $\Lambda \in \mathbb{R}^{K \times K}$ è una matrice diagonale.
Viene quindi effettuata un'associazione tra il vettore di caratteristiche non osservate del nodo i, $\mathbf{u}_i = \{u_{i,1}, \ldots, u_{i,K}\}$, e il comportamento del nodo nella rete.
Quello che accade è che la similarità tra i fattori latenti di due nodi distinti - la quale rappresenta l'equivalenza stocastica dei due attori - assieme al segno dei valori della matrice Λ, contribuisce alla natura della loro relazione. Per valori negativi sulla diagonale della matrice, infatti, ci si trova di fronte ad eterofilia, mentre per valori positivi si parla di omofilia [21].

2.2.2 Additive and Multiplicative Effects Model

Un modello in grado di combinare gli effetti additivi e moltiplicativi è l'Additive and Multiplicative Effects Model (AME) [18], il quale include nella formulazione dell'SRRM interazioni moltiplicative di caratteristiche latenti di nodo tramite l'espressione

$$y_{i,j} = \beta^\top \mathbf{x}_{i,j} + \mathbf{u}_i^\top \mathbf{v}_j + a_i + b_j + \varepsilon_{i,j} \tag{2.7}$$

dove

$$\begin{aligned}
(\mathbf{u}_1, \mathbf{v}_1), \ldots, (\mathbf{u}_n, \mathbf{v}_n) &\overset{i.i.d.}{\sim} N_{2r}(\mathbf{0}, \Psi) \\
(a_1, b_1), \ldots, (a_n, b_n) &\overset{i.i.d.}{\sim} N_2(\mathbf{0}, \Sigma) \\
\{(\varepsilon_{i,j}, \varepsilon_{j,i}) \mid i < j\} &\overset{i.i.d.}{\sim} N_2\left(\mathbf{0}, \sigma^2 \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)\right)
\end{aligned}$$

con $\mathbf{u}_i$ e $\mathbf{v}_j$ sono vettori gaussiani r-dimensionali con media nulla, i.i.d. tra i nodi con $\text{Cov}(\mathbf{u}_i, \mathbf{v}_j) = \Psi_{u,v}$ e quindi $\mathbb{E}\left[\mathbf{u}_i^\top \mathbf{v}_j \mathbf{u}_j^\top \mathbf{v}_k \mathbf{u}_k^\top \mathbf{v}_i\right] = \mathbb{E}\left[\mathbf{u}_i^\top \mathbf{v}_i\right]^3 = \text{tr}(\Psi_{u,v})^3$.
In particolare, di tratta di un AME gaussiano in quanto si assume che i dati seguano una distribuzione normale.
La sua forma matriciale risulta essere

$$\mathbf{Y} = \mathbf{M} + \mathbf{a}\mathbf{1}^\top + \mathbf{1}\mathbf{b}^\top + \mathbf{U}\mathbf{V}^\top + \mathbf{E} \tag{2.8}$$

dove $[\mathbf{M}]_{i,j} = \beta^\top \mathbf{x}_{i,j}$, $\mathbf{a} = (a_1, \ldots, a_n)$, $\mathbf{b} = (b_1, \ldots, b_n)$ e $\mathbf{U}$ e $\mathbf{V}$ sono matrici $n \times r$ con l'i-esima riga uguale a $\mathbf{u}_i$ e $\mathbf{v}_i$ rispettivamente, data r la dimensione dei vettori latenti.
Questo modello include il modello per le covarianze SRM per gli a_i, i b_j e i $\varepsilon_{i,j}$ oltre ad un modello ad effetti casuali per $\mathbf{u}_i$ e $\mathbf{v}_i$ utile per rappresentare le possibili dipendenze di terzo ordine della sociomatrice.
Da un punto di vista interpretativo l'effetto moltiplicativo $\mathbf{u}_i^\top \mathbf{v}_j$ è una media che quantifica la dipendenza di terzo ordine che può indicare l'omissione di una variabile di regressione o di una struttura di gruppo tra i nodi. Questa interpretazione si basa sull'osservazione che la presenza o l'intensità del legame è spesso collegata a similitudini a livello delle caratteristiche dei nodi. Ad esempio, se x_i è l'indicatore dell'appartenenza del nodo i ad un certo gruppo o ha una certa caratteristica, allora $x_i x_j$ indica se i nodi i e j si trovano nello stesso gruppo e questo fatto potrebbe avere un qualche effetto sulla loro relazione $y_{i,j}$.
Se effettivamente si riscontra un'associazione positiva tra $x_i x_j$ e $y_{i,j}$ allora si parla di omofilia, anti-omofilia se l'associazione è negativa. Per quantificare l'omofilia in un SRRM è sufficiente creare un regressore diadico $x_{i,j}$ combinando i regressori di nodo x_i e x_j tramite moltiplicazione o qualche altra operazione.

Tuttavia questo metodo non garantisce che tutti gli attributi di nodo rilevanti siano inclusi nei dati, questo giustifica la presenza del termine $\mathbf{u}_i^\top \mathbf{v}_j$, dove $\mathbf{u}_i$ e $\mathbf{v}_i$ rappresentano rispettivamente dei fattori latenti non osservati del nodo i come mittente e come destinatario.

Capitolo 3

I dati

I dati presi in esame consistono in tre dataset differenti. Nel seguito vengono descritte le elaborazioni che sono risultate necessarie per l'estrazione delle informazioni utili alle analisi.

3.1 Le ricette

Si dispone di una lista di 56 144 ricette provenienti dai siti americani Allrecipes e Epicurious e da quello coreano MenupAn, da cui ne sono state rimosse 354 in quanto composte da un unico ingrediente.
Per ogni ricetta vengono indicati gli ingredienti utilizzati, per un totale di 381 ingredienti distinti, e l'area geografica.

SouthernEuropean	olive_oil	garlic	bread

Figura 3.1: Esempio di riga del dataset.

In Figura 3.1 si può vedere come viene registrata la ricetta della classica bruschetta.
Il primo campo in Figura 3.1 fa riferimento all'area di provenienza della cucina a cui viene attribuito il piatto secondo i siti, secondo

Area geografica	Numero di ricette	Cucine incluse
Nord America	41 263	Americana, canadese, cajun, creola, soul food del sud e quella degli Stati Uniti sud-occidentali
Europa meridionale	4 161	Greca, italiana, mediterranea, spagnola e portoghese
America Latina	2 910	Caraibica, centroamericana, sudamericana e messicana
Europa occidentale	2 637	Francese, austriaca, belga, inglese, scozzese, olandese, svizzera, tedesca e irlandese
Asia orientale	2 484	Coreana, cinese e giapponese
Medio oriente	641	Iraniana, ebrea, libanese e turca
Asia meridionale	614	Bengalese, indiana e pachistana
Sud-est asiatico	454	Indonesiana, malese, filippina, thai e vietnamita
Europa orientale	380	Slava e russa
Africa	351	Marocchina e africana in generale
Europa settentrionale	249	Scandinava

Tabella 3.1: Il numero di ricette ed il tipo di cucina a cui fa riferimento ogni zona geografica.

il criterio presentato in Tabella 3.1, da cui si può anche notare un evidente sbilanciamento nella numerosità dei dati a favore la cucina americana, dovuto probabilmente alla scelta dei siti da cui estrarre i dati.

Si tratta di una fonte in grado di fornire informazioni sulle abitudini alimentari nelle varie culture e le combinazioni tra gli ingredienti. Nello specifico, per ogni ingrediente i si ricava il numero di ricette $N^R_{c,i}$ in cui viene utilizzato in ogni zona c, così come per ogni coppia di nodi (i, j) viene calcolato il numero di ricette $N^R_{c,i,j}$ in cui appaiono assieme in ciascuna cucina c.

Da questo primo dato viene ricavata l'autenticità [1] prima dell'ingrediente i, poi della coppia (i, j), nella cucina c, calcolate rispettivamente come

$$\mathcal{A}_{c,i} = \frac{N^R_{c,i}}{N^R_c} - \mathbb{E}_{b \neq c}\left[\frac{N^R_{b,i}}{N^R_b}\right], \tag{3.1}$$

e

$$\mathcal{A}_{c,i,j} = \frac{N^R_{c,i,j}}{N^R_c} - \mathbb{E}_{b \neq c}\left[\frac{N^R_{b,i,j}}{N^R_b}\right] \tag{3.2}$$

dove N^R_c è il numero di ricette registrate della cucina c.

Per quantificare la tendenza di una cultura alimentare ad utilizzare due ingredienti nello stesso piatto si usa l'informazione mutua puntuale [7]

$$\begin{aligned} \texttt{IMP}_c(i, j) &= \log \frac{p^c_{i,j}}{p^c_i \cdot p^c_j} \\ p^c_i = \frac{N^R_{c,i}}{N^R_c}, &\quad p^c_{i,j} = \frac{N^R_{c,i,j}}{N^R_c}. \end{aligned} \tag{3.3}$$

Tramite questa misura [22] si compara la presenza di due ingredienti nello stesso piatto con la probabilità del loro uso congiunto sotto assunzione di indipendenza.

Si tratta di una funzione simmetrica e continua su tutto l'asse reale, cioè

$$\texttt{IMP}(i,j) = \texttt{IMP}(j,i),$$
$$\texttt{IMP}(i,j) \in \mathbb{R}$$

la quale indica una maggiore propensione all'utilizzo congiunto di due ingredienti per valori crescenti. In particolare, due valori di particolare interesse sono

$$\texttt{IMP}_c(i,j) = 0,$$

per il quale la misura suggerisce indipendenza tra la coppia di nodi i e j, e

$$\texttt{IMP}_c(i,j) = -\log(p_i^c) = -\log(p_j^c),$$

in presenza della quale si dice che i e j sono perfettamente associati. Infine, per ogni ricetta viene effettuato il conteggio degli ingredienti utilizzati.

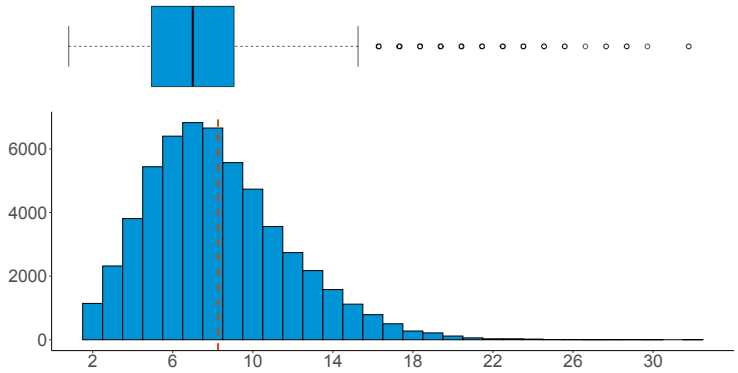

Figura 3.2: Distribuzione del numero di ingredienti per ricetta. Le quantità registrate vanno da 2 a 32, con media pari a 8.265 (linea tratteggiata) e quartili in corrispondenza dei valori 6, 8 e 10.

Come si può notare dalla Figura 3.2, per quanto i valori spazino abbastanza, la maggioranza delle ricette usa circa 8 ingredienti, tendenza che viene confermata anche effettuando una distinzione per area geografica.

3.2 Gli ingredienti

Per fornire una maggiore caratterizzazione degli ingredienti si dispone di un'attribuzione di ogni ingrediente ad una diversa categoria alimentare, per un totale di 14 categorie alimentari distinte.
L'ultimo pezzo di informazione è messo a disposizione dalla lista di adiacenza della variabile che registra il numero di composti organici condivisi da ogni coppia di ingredienti [3], quantificando quindi la similarità tra i sapori dei due nodi.

bread	olive_oil	1
bread	garlic	3

Figura 3.3: Lista di adiacenza della ricetta considerata come esempio.

Delle 72 390 possibili coppie poco più del 57% - 41 430 per l'esattezza - hanno almeno un composto organico in comune, con un massimo di 199 che sono quelli condivisi dal manzo crudo e quello cotto, ed una media di 7.959.
Oltre a fornire un'informazione sugli ingredienti, permette di caratterizzare le diverse cucine ricavando per ogni ricetta il numero medio di composti organici condivisi dai suoi ingredienti, tramite la formula [1]

$$\texttt{CO}(R) = \frac{2}{n_R(n_R-1)} \sum_{i,j \in R} x_{i,j} \tag{3.4}$$

dove n_R è il numero di ingredienti della ricetta R ed $x_{i,j}$ il numero di composti organici comuni agli ingredienti i e j. Ad esempio, gli ingredienti della bruschetta presentano mediamente 1.3334 composti organici in comune, come si può evincere dalla lista di adiacenza dei composti organici dei suoi ingredienti.

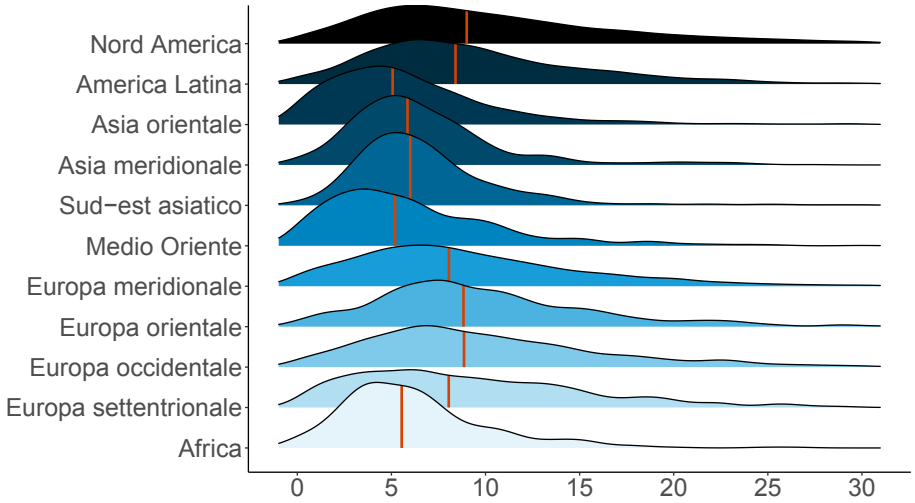

Figura 3.4: Distribuzione del numero medio di composti organici nelle ricette, in cui la linea verticale arancione indica la mediana, divisi per area geografica.

Come si può chiaramente notare dalla Figura 3.4 il numero medio di composti organici condivisi dalle ricette risulta un fattore maggiormente discriminante, essendo le cucine orientali nettamente più sbilanciate verso l'utilizzo di sapori simili nello stesso piatto, come si può notare dalla presenza di distribuzioni maggiormente distribuite lungo tutto l'asse considerato.
La decisione di procedere con l'elaborazione interpretando i nodi come vertici di una rete nasce proprio da questo dataset, in cui l'informazione è tradotta in una variabile diadica ed in quanto tale

inserita all'interno di un contesto di analisi di rete. L'approccio della network analysis nasce quindi dal desiderio di incentrare l'analisi su quello che del cibo conta di più: il sapore.

Capitolo 4

Le reti di ingredienti

Nelle ricette provenienti dall'Asia orientale vengono usati 242 ingredienti differenti, mentre la rete del Nord America presenta 354 nodi.

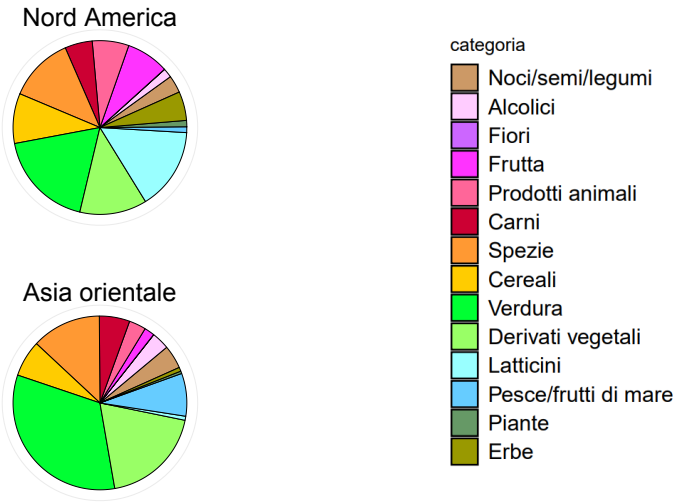

Figura 4.1: Utilizzo delle varie categorie alimentari nelle ricette.

Sin dal diverso uso delle categorie alimentari, rappresentato in Figura 4.1, si possono notare delle differenze. La cucina americana è

chiaramente caratterizzata da un maggiore uso di latticini, prodotti di origine animale e frutta, mentre quella asiatica da un largo uso di pesce e verdure.

4.1 I composti organici

Attraverso l'analisi dei grafi e altri strumenti ad hoc per l'interpretazione e la raffigurazione dei dati di rete, risulta possibile cogliere quelli che sembrano essere degli ulteriori criteri ricorrenti nella composizione dei piatti e nella scelta degli elementi da inserirvi.

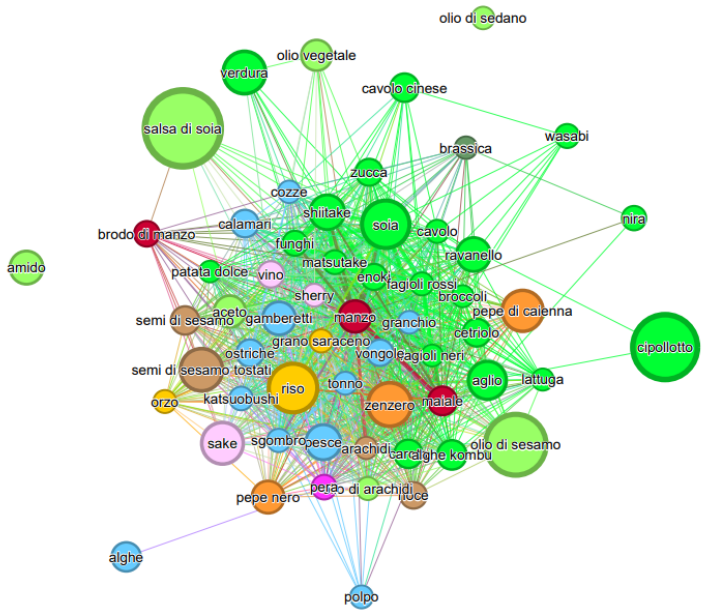

Figura 4.2: Rete dei composti organici condivisi tra gli ingredienti autentici usati almeno nell'1% delle ricette dell'Asia orientale.

Osservando la rete di composti organici condivisi dagli ingredienti autentici delle ricette asiatiche, rappresentata in Figura 4.2, risultano già evidenti alcune caratteristiche.

Oltre ad una netta presenza di verdure, che compongono quasi il 40% dei nodi, derivati vegetali e pesce, si nota sin dal primo sguardo come la rete non sia troppo compatta e come all'aumentare dell'autenticità aumenti la tendenza del nodo a posizionarsi in prossimità degli estremi del grafo. Infatti, la disposizione dei nodi dipende dalla loro forza, un valore che spazia da 1 a 6 864 con una media di 2 041.1, mentre la loro dimensione rispecchia l'autenticità registrata.

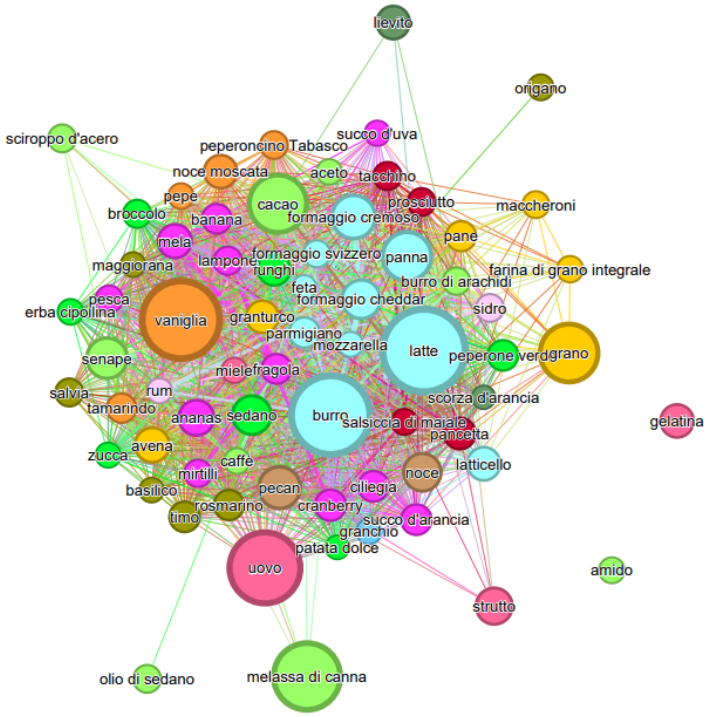

Figura 4.3: Rete dei composti organici condivisi tra gli ingredienti autentici usati almeno nell'1% delle ricette del Nord America.

Dei 59 nodi considerati per la rappresentazione della rete di composti organici degli ingredienti asiatici, solo 8 si trovano anche tra i 69 ingredienti autentici del Nord America, elementi della rete presentata in Figura 4.3. Oltre a presentare una maggiore varietà nelle

categorie alimentari, presenta un aspetto più compatto, in cui ad un maggiore grado di autenticità sembra corrispondere anche un alto valore della forza. Fatto ulteriormente marcato dalla presenta di valori generalmente più alti nella forza dei composti organici americani: i valori spaziano da 1 fino ad un valore di 9 803, con una media pari a 2 891.

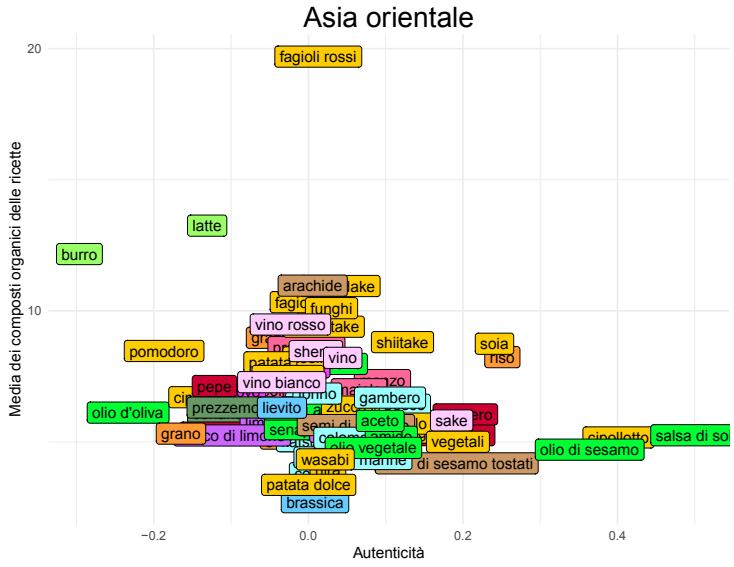

Figura 4.4: Rapporto tra autenticità e numero medio di composti organici nelle ricette, per gli ingredienti usati in più dell'1% dei piatti.

Elementi che caratterizzano entrambi i grafi sono invece, il valore dei coefficienti di clustering pesati e le densità. Gli indici descrittivi della tendenza alla formazione di triadi registrano valori di $C^{+} = 0.94$ e $C^{\times} = 0.04$, suggerendo che le due reti abbiano comunque dei comportamenti abbastanza simili una volta considerato solo il valore degli archi, con una forte propensione a creare triadi ma con valori complessivi bassi. Per quanto riguarda le densità, invece, entrambe

le reti riportano un valore di 0.6, che rimane costante anche per il sotto-grafo usato per la rappresentazione nel caso del continente asiatico, mentre aumenta leggermente, fino a raggiungere il 70% degli archi ammissibili, nel caso della rete ristretta agli ingredienti caratteristici della cucina nord americana.
Per quanto riguarda però il rapporto tra autenticità e l'abbinamento dei sapori, un discorso simile è riproponibile osservando le ricette in cui vengono usati gli ingredienti.

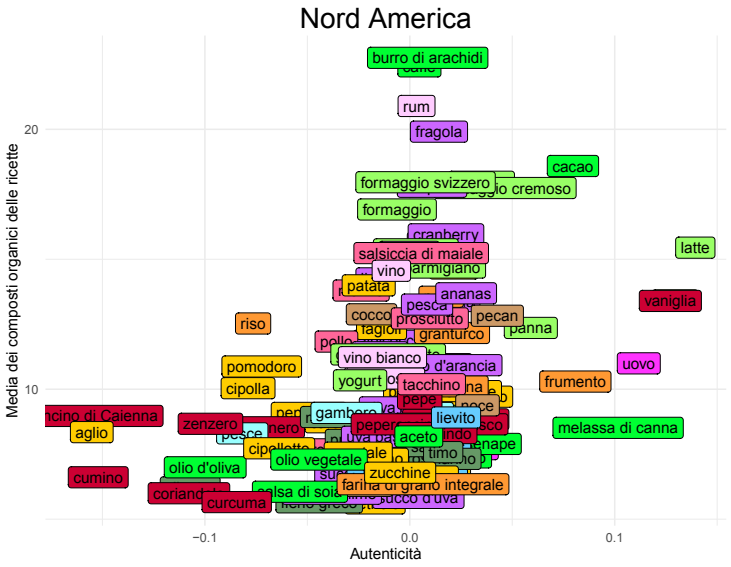

Figura 4.5: Rapporto tra autenticità e numero medio di composti organici nelle ricette, per gli ingredienti usati in più dell'1% dei piatti.

Come si può notare facendo riferimento alla Figura 4.4 il numero medio di composti organici condivisi nelle ricette in cui appare un ingrediente è tendenzialmente basso in Asia orientale, ma soprattutto questo valore presenta un andamento decrescente all'aumentare dell'autenticità dell'ingrediente.

Contrariamente, il Nord America presenta un numero medio di composti organici decisamente più altro, in cui questa propensione viene confermata dagli elementi caratteristici.

4.2 L'informazione mutua puntuale

Per quanto riguarda le reti che registrano l'informazione mutua puntuale degli ingredienti, cioè la misura che indica la propensione di due ingredienti ad essere usati assieme, in caso di valore positivo, e la tendenza ad evitare il loro accostamento, per i valori negativi, ci si trova di fronte ad una densità nettamente inferiore.
La rete che fa riferimento ai dati dell'Asia orientale registra solo il 24% delle scelte di utilizzare insieme due ingredienti come non casuali, mentre nelle ricette americane gli ingredienti sembrano essere maggiormente interdipendenti nel loro utilizzo, con una densità pari a 0.41.
In Figura 4.6 è riportata la rete che registra gli ingredienti che presentano una maggiore propensione ad essere utilizzati contemporaneamente. La dimensione dei nodi è data dal numero di composti organici viene mediamente condivisa all'interno delle ricette che prevedono l'utilizzo dell'ingrediente in questione, infatti le dimensioni dei nodi sono principalmente contenute. Si noti, inoltre, come siano verdure e pesce a popolare la zona maggiormente densa di collegamenti, ad indicare come la scelta di un loro utilizzo combinato sia guidato da precise preferenze.
Diversamente, altri ingredienti come cereali e prodotti di origine vegetale, quali il riso e la salsa di soia, per quanto fortemente autentici della cucina considerata, presentano pochi collegamenti all'interno della rete.

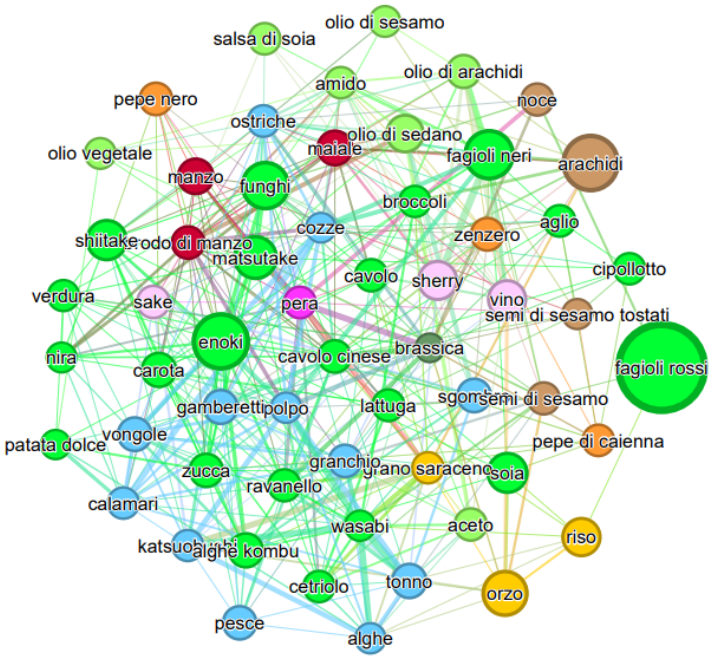

Figura 4.6: Rete dell'informazione mutua puntuale positiva tra gli ingredienti autentici usati almeno nell'1% delle ricette dell'Asia orientale (è rappresentata solo la metà maggiormente significativa di archi per facilitare la lettura del grafo).

Questo comportamento prova una grande varietà negli abbinamenti effettuati, in quando, in caso contrario, vista l'elevata presenza di questi ingredienti all'interno delle ricette, sarebbero presenti molti più archi a partire da questi nodi. Infatti, il riso viene usato ben 843 volte, un numero notevole considerando che il terzo quantile della distribuzione del numero di utilizzi degli ingredienti ha il terzo quantile in corrispondenza dei 51 utilizzi, mentre la salsa di soia è l'ingrediente più usato in assoluto con 1358 utilizzi.
Per quanto riguarda invece la rete della propensione ad evitare una certa coppia, non risultano evidenti pattern particolari, i vari ingre-

dienti rifiutano l'abbinamento senza evidenti criteri attraverso una prima analisi.

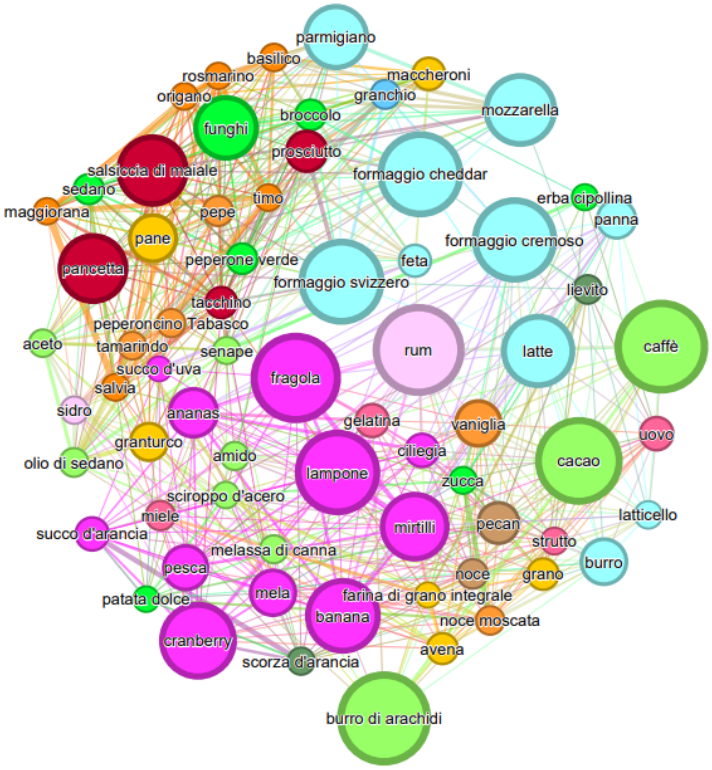

Figura 4.7: Rete dell'informazione mutua puntuale positiva tra gli ingredienti autentici usati almeno nell'1% delle ricette del Nord America (è rappresentata solo la metà maggiormente significativa di archi per facilitare la lettura del grafo).

In generale la rete dell'Asia orientale presenta maggiori informazioni per quanto riguarda gli abbinamenti in positivo, riportando forza maggiore di zero, sin dal primo quantile della sua distribuzione, il quale viene registrato pari a 35.72. Tuttavia, sia la rete completa che le due sotto-reti delle forze attrattive e repulsive, presentano una bassa propensione alla formazione di triadi, arrivando ad avere

$C^{\times} = 0$ nel caso repulsivo.
La rete delle relazioni attrattive ottenuta dagli ingredienti autentici delle ricette nord americane, presentata in Figura 4.7, evidenzia quasi un comportamento a cluster, che tuttavia si è dimostrato non significativo. Si può notare infatti come gli ingredienti delle stesse categorie tendano ad attrarsi e a concentrarsi nella stessa area, tuttavia, come succede per l'altra area considerata, entrambi i coefficienti di clustering pesati riportano valori bassi, con $C^{\times} = 0.05$.

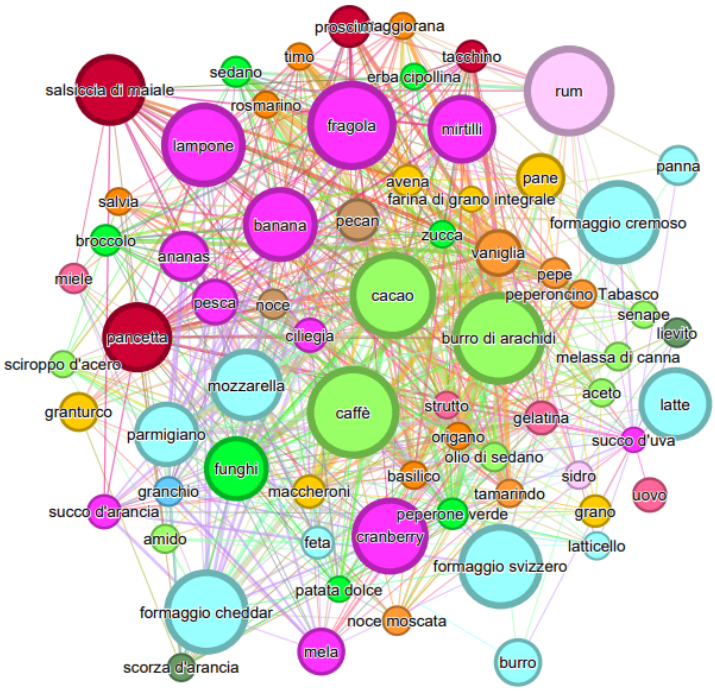

Figura 4.8: Rete dell'informazione mutua puntuale negativa tra gli ingredienti autentici usati almeno nell'1% delle ricette del Nord America (è rappresentata solo la metà maggiormente significativa di archi per facilitare la lettura del grafo).

Contrariamente, la forza riporta valori più estremi di quanto fatto

in precedenza, con una media pari a 115.7, contro il 78.1 del caso asiatico.
Questo suggerisce una certa consistenza nella scelta degli ingredienti nello stesso piatto, tendenza che sembra essere confermata anche dalla Figura 4.8, in cui appare la rete complementare a quest'ultima. Si vede come in questo caso la distribuzione dei nodi è molto meno legata alla categoria, in quanto la maggior parte degli archi collegano ingredienti di categorie alimentari differenti. La propensione a formare triadi continua ad essere bassa, con $C^{\times} = 0$, ma non come in precedenza, in quanto $C^{+} = 0.43$, contro lo 0.29 della rete di repulsione degli ingredienti asiatici, e la forza raggiunge valori assoluti superiori, con una media di 92, diversamente dal caso precedente in cui la media della distribuzione delle forze era solo 18.1.

Capitolo 5

La modellazione della rete

Si applicano i modelli per dati di rete alla rete di informazione mutua puntuale, con l'obiettivo di confermare o smentire quanto emerso dalle analisi esplorative. Le conclusioni raggiunte finora sono che, nella regione del Nord America, ad un aumento dei composti organici condivisi coincide un aumento della probabilità della coppia di apparire nelle ricette, tendenza opposta a quella riscontrata nella cucina dell'Asia orientale in cui sembrano essere favoriti piatti con sapori tra loro anche molto distanti.

Le stime vengono ricavate tramite i comandi dei pacchetti **amen** del software **R**, i quali vengono usati per modellare l'effetto delle variabili esplicative diadiche e di nodo, descritte in Tabella 5.1, nella determinazione del valore dell'informazione mutua puntuale tra le coppie di ingredienti.

I risultati del modello SRM e dell'Eigenmodel, della libreria **eigenmodel**, non vengono presentati in quanto non in grado di fornire contenuti significativi ai fini della domanda di ricerca.

Nome	Tipologia	Significato
Y	diadica	Informazione mutua puntuale delle coppie di ingredienti, variabile relazionale oggetto della modellazione
categ.node	di nodo	Categoria alimentare associata all'ingrediente
c.org.ric.node	di nodo	Media del numero di composti organici mediamente condivisi dagli ingredienti delle ricette in cui appare il nodo
autent.node	di nodo	Autenticità registrata
dim.ric.node	di nodo	Numero di ingredienti mediamente previsti dalle ricette in cui viene usato l'ingrediente
sapori.dyad	diadica	Numero di comporti organici condivisi dalla coppia di ingredienti, i quali svolgono un ruolo centrale nella determinazione del loro sapore
c.org.ric.dyad	diadica	Media del numero di composti organici mediamente condivisi dagli ingredienti dei piatti in cui la diade appare contemporaneamente
autent.dyad	diadica	Autenticità della coppia nella cucina dell'area considerata
dim.ric.dyad	diadica	Numero di ingredienti mediamente previsti dalle ricette in cui la coppia di ingredienti viene usata

Tabella 5.1: Variabili utilizzate per la stima dei modelli.

5.1 Social Relation Regression Model

Essendo l'obiettivo quello di valutare la relazione tra l'informazione mutua puntuale e le altre variabili diadiche e di nodo, si stima un modello SRRM che, essendo la matrice di adiacenza simmetrica, diventa [23]

$$y_{i,j} = \beta^\top \mathbf{x}_{i,j} + a_i + a_j + \varepsilon_{i,j} \tag{5.1}$$

con

$$a_1, \ldots, a_n \overset{i.i.d.}{\sim} N(0, \sigma_a^2)$$
$$\{(\varepsilon_{i,j}, \varepsilon_{j,i}) \mid i \neq j\} \overset{i.i.d.}{\sim} N(0, \sigma_\varepsilon^2)$$

dove $\mathbf{x}_{i,j}$ è il vettore p-dimensionale di regressori diadici e di nodo.
Il risultato sono le stime dei coefficienti, di cui vengono riportate deviazione standard, lo *z-score* e il *p-value*, riportate in Tabella 5.2 per la rete delle ricette asiatiche.
Si nota una forte influenza delle variabili diadiche che descrivono valori in riferimento al complessivo di ricette in cui viene usata la coppia, entrambe con un effetto positivo sulla variabile risposta.
Questo risultato, in riferimento alle variabili che registrano l'omogeneità di sapori, appare in contraddizione con la tendenza della cucina asiatica a produrre piatti con sapori distanti tra loro, ma il valore negativo dell'autenticità, sia livello diadico che di nodo, suggerisce un aspetto interessante: l'utilizzo combinato di ingredienti autentici, a causa del loro alto utilizzo nelle ricette della cucina interessata, non suggerisce una dipendenza tra i due.
Lo scarso numero di composi organici condivisi nelle ricette, quindi, non risulta attribuibile a una generale tendenza ad abbinare ingredienti dai sapori molto diversi, ma solo all'uso diffuso di relativamente pochi ingredienti caratterizzati da composti organici poco comuni.

	pmean	psd	z-stat	p-val
intercept	-0.073	0.200	-0.366	0.714
categ.node	-0.006	0.005	-1.205	0.228
c.org.ric.node	0.010	0.005	1.954	0.051
autent.node	-2.714	0.260	-10.449	0.000
dim.ric.node	-0.005	0.007	-0.711	0.477
sapori.dyad	0.000	0.000	-0.917	0.359
c.org.ric.dyad	0.120	0.002	66.467	0.000
autent.dyad	-1.118	0.399	-2.802	0.005
dim.ric.dyad	0.067	0.001	48.744	0.000

	pmean	psd
va	0.088	0.009
ve	0.390	0.003

Tabella 5.2: Parametri del modello SRRM applicato alla rete dell'Asia orientale.

Gli ingredienti autentici, a causa della loro elevata presenza all'interno delle ricette, causano un mascheramento delle tendenze più generali in una prima fase delle analisi.
Risultano non essere significative solo la categoria alimentare, il numero di composti organici condivisi dalle coppie e la numerosità media degli ingredienti nelle ricette in cui è previsto l'uso del nodo, mentre l'omogeneità nei sapori nelle ricette che coinvolgono il nodo hanno un effetto positivo che risulta però essere basso e al limite del significativo, diversamente da quello che accade nella stima dello stesso modello applicato ai dati delle ricette nord americane, il cui risultato è riportato in Tabella 5.3. La media dei composti organici condivisi nelle ricette che prevedono l'utilizzo dell'ingrediente,

infatti, registra un effetto negativo e significativo.
Un'altra discrepanza importante è data dal comportamento della variabile diadica di autenticità, la quali risulta ora non significativa, mentre aumenta il valore esplicativo del numero di composti organici dalle coppie, che riporta però un coefficiente nullo.
Sono le variabili diadiche che registrano gli effetti dei composti organici medi nelle ricette e la loro dimensione in termini di ingredienti utilizzati a riconfermarsi le variabili di maggiore effetto sulla variabile risposta, con coefficienti positivi anche se i valori sono più bassi rispetto a quelli registrati nel modello precedente.

	pmean	psd	z-stat	p-val
intercept	0.187	0.235	0.794	0.427
categ.node	-0.007	0.005	-1.292	0.196
c.org.ric.node	-0.022	0.005	-4.786	0.000
autent.node	-2.292	0.690	-3.323	0.001
dim.ric.node	0.006	0.010	0.631	0.528
sapori.dyad	0.002	0.000	5.069	0.000
c.org.ric.dyad	0.040	0.001	41.903	0.000
autent.dyad	1.869	1.101	1.697	0.090
dim.ric.dyad	0.033	0.001	34.917	0.000

	pmean	psd
va	0.129	0.010
ve	0.602	0.003

Tabella 5.3: Parametri del modello SRRM applicato alla rete del Nord America.

Questa formulazione per la descrizione dell'informazione mutua puntuale presenta una buona capacità di adattamento ai dati, soprat-

tutto nel secondo caso, tuttavia presenta dei limiti nella sua capacità di catturare eventuali dipendenze triadiche. Pertanto si procede utilizzando l'implementazione della formulazione AME del modello per provare a considerare anche questo aspetto.

5.2 Additive and Multiplicative Effects Model

La stima del modello AME permette di considerare sia le variabili osservate che gli effetti latenti tra le esplicative, andando così a cogliere anche gli effetti triadici, quali possono essere la transitività e la ciclicità - tendenza a formare triadi ordinate - che nel caso di matrice di adiacenza simmetrica coincidono.

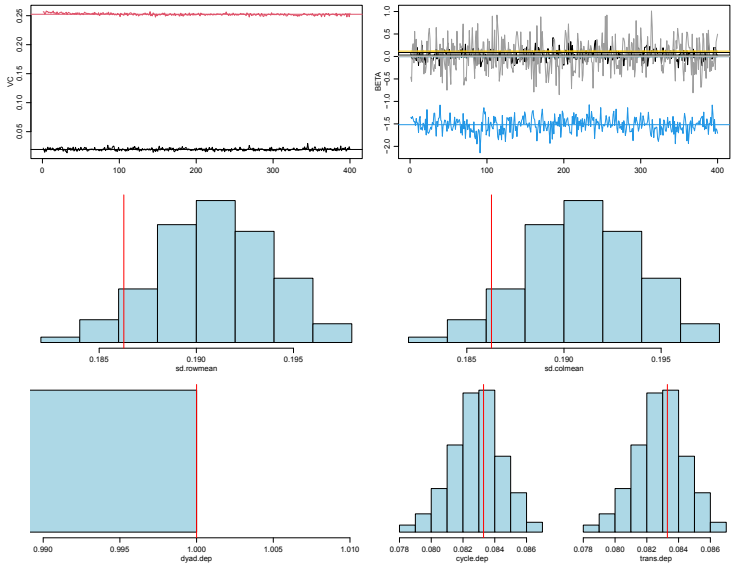

Figura 5.1: Bontà di adattamento del modello AME con nove fattori latenti stimato sulla rete dell'Asia orientale.

Nell'adattamento di questo modello ai dati provenienti dal dataset dell'Asia orientale si è notato un miglioramento della bontà di adattamento, soprattutto in termini di dipendenze triadiche, all'aumentare del numero di fattori latenti considerati.
Tale miglioramento raggiunge un massimo in corrispondenza del modello AME con otto fattori latenti fornendo un ottimo adattamento, soprattutto a livello di dipendenze triadiche. Questa bontà nelle stime viene riportata in Figura 5.1, in cui la prima riga presenta l'andamento delle stime che corrisponde alle realizzazioni delle catene di Markov tramite cui viene effettuato l'adattamento, la prima per le varianze, la seconda per i coefficienti del modello. Gli istogrammi sottostanti riportano, invece, le distribuzioni a posteriori delle statistiche di rete, confrontate con il valore delle statistiche osservate, indicate dalle linee rossa. Essendo in un contesto di grafo indiretto gli effetti di riga e colonna risultano uguali, così come le dipendenze triadiche, mentre la dipendenza diadica è pari ad 1.
I risultati in termini di significatività dei coefficienti stimati, riportati in Tabella 5.4, risultano abbastanza coerenti. Si riconfermano altamente significative le variabili diadiche del numero medio di composti organici condivisi nelle ricette e il numero di ingredienti necessari per la composizione del piatto, entrambe con effetti positivi, mentre l'autenticità della diade perde di significatività.
A livello di nodo il coefficiente dell'autenticità diminuisce di valore in modulo, ma rimane significativo e negativo, mentre categoria e composti organici nelle ricette risultano non apportare sufficiente informazione sulla variabile risposta.
Di particolare interesse risulta essere l'elevato numero di fattori latenti, i cui valori dei coefficienti stimati, riportati in Tabella 5.5, suggeriscono la presenza di omofilia tra i nodi [21].

	pmean	psd	z-stat	p-val
intercept	0.031	0.112	0.278	0.781
categ.node	-0.003	0.003	-1.118	0.263
c.org.ric.node	0.004	0.003	1.540	0.124
autent.node	-1.523	0.166	-9.147	0.000
dim.ric.node	-0.004	0.004	-1.038	0.299
sapori.dyad	0.000	0.000	-0.640	0.522
c.org.ric.dyad	0.119	0.002	67.622	0.000
autent.dyad	0.027	0.346	0.079	0.937
dim.ric.dyad	0.088	0.001	66.934	0.000

	pmean	psd
va	0.019	0.002
ve	0.252	0.002

Tabella 5.4: Parametri del modello AME con otto fattori latenti applicato alla rete dell'Asia orientale.

Tuttavia, la mancata conoscenza di quali caratteristiche di nodo vengano colte da tali fattori rende complesso trarre delle ulteriori conclusioni sulle caratteristiche che rendono le scelte di inserire due ingredienti nello stesso piatto non indipendenti.

-123.81	25.84	27.64	28.95	33.13	36.88	39.48	76.08

Tabella 5.5: Stima dei coefficienti degli effetti latenti del modello AME stimato sulla rete dell'Asia orientale.

Un discorso diverso può essere fatto per il modello adattato ai dati nord americani, il quale, con soli due effetti latenti presenta un ottimo adattamento, come si può notare in Figura 5.2.

La variabile maggiormente significativa è la numerosità degli ingredienti nelle ricette in cui viene previsto l'uso della diade, il cui coefficiente positivo sta ad indicare come l'aumentare del numero di ingredienti necessari renda maggiormente ragionevole aspettarsi che i due nodi vengano usati assieme.

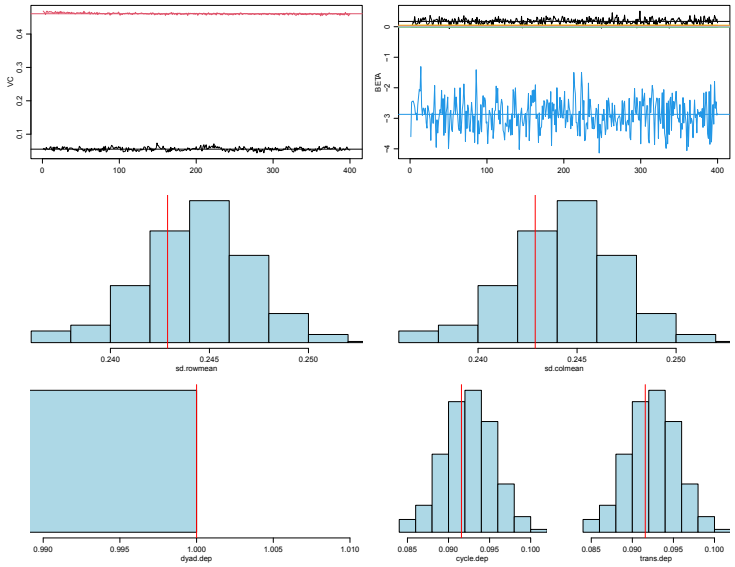

Figura 5.2: Bontà di adattamento del modello AME con due fattori latenti stimato sulla rete del Nord America.

Guardando gli altri coefficienti, riportati nella Tabella 5.6, si nota una maggiore significatività di questi ultimi rispetto a quanto accade per la rete asiatica. Tuttavia, risulta interessante notare come tutti gli effetti significativi in entrambi condividano ordini di grandezza, segno e significatività, evidenziano un comportamento comune all'interno delle due cucine.
L'unica differenza tra gli effetti delle variabili osservate delle due aree è la presenza di un effetto negativo del numero di composti organici mediamente condivisi nelle ricette in cui è presente il nodo.

Questo fatto implica che un ingrediente usualmente usato in ricette con tanti composti organici condivisi porta ad un inferiore impatto positivo sulla ricetta.

	pmean	psd	z-stat	p-val
intercept	0.015	0.168	0.087	0.931
categ.node	-0.005	0.004	-1.314	0.189
c.org.ric.node	-0.011	0.003	-3.370	0.001
autent.node	-2.803	0.482	-5.818	0.000
dim.ric.node	0.008	0.007	1.216	0.224
sapori.dyad	0.001	0.000	5.151	0.000
c.org.ric.dyad	0.033	0.001	36.961	0.000
autent.dyad	1.643	0.921	1.784	0.074
dim.ric.dyad	0.050	0.001	56.511	0.000

	pmean	psd
va	0.055	0.005
ve	0.461	0.002

Tabella 5.6: Parametri del modello AME con due fattori latenti applicato alla rete del Nord America.

Per quanto le stime risultino accurate risulta necessario tenere conto che a livello interpretativo ci sono dei limiti imputabili alla presenza di fattori latenti con effetti importanti, come si può notare dalla Tabella 5.7.

-170.07	152.54

Tabella 5.7: Stima dei coefficienti degli effetti latenti del modello AME stimato sulla rete del Nord America.

Capitolo 6

Conclusioni

La branca di ricerca detta gastronomia computazionale [5] si pone una serie di domande come "Perché mangiamo quello che mangiamo?" o "Da cosa dipende il sapore di un ingrediente?". In questo frangente si prova a rispondere al seguente quesito: è vera la tendenza di un abbinamento di ingredienti, composto da elementi dal gusto simile, ad avere un sapore migliore? [2]
Per farlo sono stati analizzati i dati ottenuti da dei siti americani ed asiatici di ricette, da cui sono state estratti la cultura a cui viene attribuita la rispettiva creazione e gli ingredienti necessari per la composizione di ogni piatto[1]. Di ogni combinazione possibile di questi ultimi, inoltre, viene considerato il numero di composti organici comuni ad entrambi [3] per quantificare la similarità tra i sapori.
Tali dati sono stati elaborati sotto forma di dati di rete, tramite indici e modelli in grado di cogliere le dipendenze e le complesse relazioni intrinseche nella natura di questo genere di dato [8], per spiegare come le informazioni sulle ricette e le caratteristiche degli ingredienti presenti nelle varie cucine influiscano sulla predisposizione di una coppia di ingredienti ad essere usata assieme, ponendo l'accento sulle

similarità e differenze tra i sapori che vengono combinati.
Le analisi sono state incentrare su due aree geografiche ben precise: il Nord America e l'Asia orientale. Tale scelta è stata dettata sia dalla volontà di effettuare un confronto con l'articolo da cui è nato questo lavoro [1], sia dalla necessità di tenere in considerazione la fonte dei dati. Attenendosi alle regioni di provenienza dei siti di ricette si suppone di disporre di ricette originali e senza contaminazioni da parte di altre culture culinarie.
L'analisi descrittiva, effettuata con gli indici propri dell'analisi di rete, individua le stesse evidenze dell'articolo originario: la cucina occidentale sembra rispondere in maniera affermativa alla domanda di ricerca, mentre la cucina orientale sembra essere in disaccordo. Tuttavia, tramite la stima di modelli AME [17], selezionati in quanto in grado di prendere in considerazione sia variabili esplicative che la presenza di fattori latenti, sono stati raggiunti dei risultati simili per entrambe le aree: la propensione di due ingredienti ad apparire assieme in una preparazione aumenta all'aumentare della media dei composti organici mediamente condivisi, sia a livello diadico che di nodo, e del numero di ingredienti medi nelle ricette in cui sono presenti entrambi, mentre diminuisce quanto più gli ingredienti sono caratteristici di quella specifica cucina rispetto a tutte le altre.
Analizzando i risultati si ottiene che un aumento della significatività nell'uso combinato di una coppia dipende dall'aumento del numero di composti organici mediamente condivisi dalle ricette in cui vengono utilizzati solitamente, variabile che risulta essere maggiormente significativa tra tutte le considerate.
Quello che si conclude è che l'ipotesi iniziale sembra essere corretta, fornendo delle indicazione per la creazione dei piatti.
Le tendenze apparentemente discordi che caratterizzano la prima fa-

se di analisi, le stesse evidenziate nello studio originale, sono un effetto confondente dato dall'ampio utilizzo degli ingredienti caratteristici, i quali presentano effettivamente un comportamento differente nelle diverse cucine.
L'ipotesi che sostiene che l'abbinamento di sapori simili porti ad un risultato tendenzialmente migliore è sostenuta anche all'interno della cucina saudita [24], completamente esterna a questo studio, a supportare l'idea che si tratti di un comportamento che trascende il gusto personale o regionale.
Tuttavia, la presenza di fattori latenti - otto nel caso della rete degli ingredienti della cucina asiatica - rende difficile considerare conclusive le spiegazioni fornite dai dati a disposizione.
Il ramo della gastronomia computazionale che studia i sapori, infatti, è un ambito di studio particolarmente complesso a causa della presenza di preferenze personali che possono far percepire lo stesso sapore in maniera anche molti diversa. Senza contare gli aspetti nutrizionali [25], i metodi di cottura [26] e le quantità dei vari ingredienti, aspetti che sarebbe sicuramente vantaggioso quantificare per spiegare almeno parte degli effetti latenti.
In presenza di tali dati sarebbe interessante riproporre un'analisi simile, vista l'adeguatezza dimostrata dagli strumenti usati, estendendo anche l'oggetto di ricerca a quesiti più specifici come l'individuazione dei motivi per cui le varie cucine regionali hanno sviluppato i gusti per cui sono oggi conosciute.

Bibliografia

(*1*) Ahn, Y.-Y., Ahnert, S., Bagrow, J. e Barabási, A.-L. (2011). Flavor network and the principles of food pairing. *Scientific Reports*, DOI: 10.1038/srep00196.

(*2*) Blumenthal, H., *The Fat Duck Cookbook*; Bloomsbury USA: 2009.

(*3*) Burdock, G., *Fenaroli's Handbook of Flavor Ingredients*; CRC Press: 2004.

(*4*) Breslin, P. e Beauchamp, G. (1995). Suppression of bitterness by sodium: variation among bitter taste stimuli. *Chemical senses 20*, 609–623.

(*5*) Goel, M. e Bagler, G. (2022). Computational gastronomy: A data science approach to food. *Journal of Biosciences 47*, DOI: 10.1007/s12038-021-00248-1.

(*6*) Sharma, T., Upadhyay, U., Kalra, J., Arora, S., Ahmad, S., Aggarwal, B. e Bagler, G. in *2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW)*, 2020, pp. 98–104.

(*7*) Teng, C.-Y., Lin, Y.-R. e Adamic, L. A. (2012). Recipe recommendation using ingredient networks. *Proceedings of the 4th annual ACM web science conference*, 298–307.

(*8*) Minhas, S., Dorff, C., Gallop, M. B., Foster, M., Liu, H., Tellez, J. e Ward, M. D. (2022). Taking dyads seriously. *Political Science Research and Methods 10*, 703–721.

(*9*) Hollstein, B., Pescosolido, B. e Smith, E. B. in *Personal Networks: Classic Readings and New Directions in Egocentric Analysis*, Small, M. L. e Perry, B. L., cur.; Structural Analysis in the Social Sciences; Cambridge University Press: 2021, pp. 44–59.

(*10*) Segre, S. (2004). A Durkheimian Network Theory. *Journal of Classical Sociology 4*, 215–235.

(*11*) Piselli, F. (1994). Famiglia e networks sociali. Tradizioni di studio a confronto. *Meridiana*, 45–92.

(*12*) Watts, D. J., *Six degrees: The science of a connected age*; WW Norton & Company: 2004.

(*13*) Fraiberger, S. P., Sinatra, R., Resch, M., Riedl, C. e Barabási, A.-L. (2018). Quantifying reputation and success in art. *Science 362*, 825–829.

(*14*) Laber, M., Klimek, P., Bruckner, M., Yang, L. e Thurner, S. (2022). Shock propagation in international multilayer food-production network determines global food availability: the case of the Ukraine war. *arXiv preprint arXiv:2210.01846*.

(*15*) Antoniou, I. e Tsompa, E. (2008). Statistical Analysis of Weighted Networks. *Discrete Dynamics in Nature and Society 2008*, DOI: 10.1155/2008/375452.

(*16*) Onnela, J.-P., Saramäki, J., Kertész, J. e Kaski, K. (2005). Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E 71*, 065103.

(*17*) Minhas, S., Hoff, P. D. e Ward, M. D. (2019). Inferential Approaches for Network Analysis: AMEN for Latent Factor Models. *Political Analysis 27*, 208–222.

(*18*) Hoff, P. (2021). Additive and Multiplicative Effects Network Models. *Statistical Science 36*, 34–50.

(*19*) Sosa, J. e Buitrago, L. (2021). A review of latent space models for social networks. *Revista Colombiana de Estadıstica 44*, 171–200.

(*20*) Hoff, P. (2007). Modeling homophily and stochastic equivalence in symmetric relational data. *Advances in neural information processing systems 20*.

(*21*) Hoff, P. D. (2013). Bayesian analysis of matrix data with rstiefel. *arXiv preprint arXiv:1304.3673*.

(*22*) Fano, R., *Transmission of Information: A Statistical Theory of Communication*; MIT Press Classics; MIT Press: 1961.

(*23*) Hoff, P. D. (2015). Dyadic data analysis with amen. *arXiv preprint arXiv:1506.08237*.

(*24*) Al-Razgan, M., Tallab, S. e Alfakih, T. (2021). Exploring the Food Pairing Hypothesis in Saudi Cuisine Using Genetic Algorithm. *Mathematical Problems in Engineering 2021*, 1–16.

(*25*) Potter, N. N. e Hotchkiss, J. H., *Food science*; Springer Science & Business Media: 2012.

(*26*) Nyati, U., Rawat, S., Gupta, D., Aggrawal, N. e Arora, A. (2019). Characterize ingredient network for recipe suggestion. *International Journal of Information Technology*, 1–8.

Appendice A

Il codice

Caricamento dei dati, la loro pulizia e la creazione delle variabili.

```
# librerie
library(tidyverse)
library(ggplot2)
library(forcats)
library(ggridges)
library(igraph)
library(DirectedClustering)
library(eigenmodel)
library(amen)

# caricato il dataset delle ricette
ricette.ingr <- ricette %>% select(- origine)

# numero di ingredienti per ricetta
dim.ric <- as.data.frame(matrix(0,
            nrow = nrow(ricette.ingr), ncol = 2))
colnames(dim.ric) <- c("origine", "n.ingr")
dim.ric$origine <- ricette$origine
```

```
for(i in 1:nrow(ricette.ingr)){
  for(j in 1:ncol(ricette.ingr)) {
    if(ricette.ingr[i,j] != "")
    {dim.ric[i,2] <- dim.ric[i,2]+1}
  }
  cat(i, "\n")
}

# esclusione delle ricette con un solo ingrediente
ricette.ingr <- ricette.ingr[dim.ric$n.ingr > 1, ]
ricette <- ricette[dim.ric$n.ingr > 1, ]
num.ingr.reg <- dim.ric[dim.ric$n.ingr > 1, ]
num.ingr.reg <- data.frame(ricette$origine, num.ingr)

# grafico sulla numerosità degli ingredienti
# nelle ricette
boxplot(num.ingr, horizontal = T, ylim = c(2.5, 32),
        xaxt = "n", col = "#0093D5", frame = F)

num.ingr.reg %>% ggplot(aes(x = num.ingr)) +
  geom_histogram(color = "black", fill = "#0093D5",
                 binwidth = 1) +
  theme_classic() +
  theme(axis.text.x = element_text(size=20),
        axis.text.y = element_text(size=20)) +
  geom_vline(aes(xintercept=mean(num.ingr)),
             color="#D54200",
             linetype="dashed",
```

```
            size=1)+
  scale_x_continuous(name = NULL, breaks=seq(2,32,4),
                     c(2,32)) +
  scale_y_continuous(name = NULL)

# caricato il dataset contenente la lista di
# adiacenza contenente il numero di composti
# organici condivisi dalle coppie con il nome
# sapori
ingr.unici <- ricette.ingr %>% as.matrix(.) %>%
                as.vector(.) %>% unique(.)
ingr.unici <- ingr.unici[- which(ingr.unici == "")]

# creazione della rete dell'uso degli ingredienti
c <- 0
ingredientinew <- matrix(NA, nrow = length(ingr.unici)*
                           (length(ingr.unici) + 1)/2,
                         ncol = 3)

for(i in ingr.unici) {
  ingr1 <- ricette.ingr %>% filter_all(any_vars(. %in% i))
  cat("ingrediente", c <- c+1, ":", i, "\n")
  b <- 1
  for(j in ingr.unici) {
    if(b <= c) {
      saporinew[((c*(c+1))/2)+b-c, 1] <- i
      saporinew[((c*(c+1))/2)+b-c, 2] <- j
      saporinew[((c*(c+1))/2)+b-c, 3] <- length(which(
        as.vector(t(ingr1)) == j))
```

```
      b <- b+1
    }
  }
}


ingredienti <- as.data.frame(ingredientinew)
colnames(ingredienti) <- c("ingrediente1",
                           "ingrediente2", "uso")

# quante volte appaiono assieme
mtx.ingr <- as.matrix(get.adjacency(graph.data.frame(
  ingredienti, directed = F), attr = "uso"))

# quanti sapori condividono
mtx.sap <- as.matrix(get.adjacency(graph.data.frame(
  sapori, directed = F), attr = "comp_org"))
mtx.sap <- mtx.sap[colnames(mtx.sap) %in% ingr.unici,
                   colnames(mtx.sap) %in% ingr.unici]

sapori <- as.matrix.data.frame(sapori, rownames.force = T)
rete.sap <- graph.adjacency(sapori, mode = "undirected",
                            weighted = T, diag = F)
sapori.add <- get.data.frame(rete.sap)
summary(sapori.add)

# utilizzo
utilizzo <- diag(mtx.ingr)
sum(utilizzo == 0)
```

```
sum(utilizzo != 0)

# utilizzo per aree geografiche
c <- 0
cucina <- matrix(NA, nrow = length(ingr), ncol = 12)
colnames(cucina) <- c("ingrediente",
                      unique(ricette$origine))
regioni <- colnames(cucina)[-1]
cucina[, 1] <- ingr.unici

for(i in ingr) {
  cat("ingrediente", c <- c+1, ":", i, "\n")
  ingr1 <- ricette %>% filter_all(any_vars(. %in% i))
  orig <- ingr1$origine
  n <- length(orig)
  for(j in regioni) {
    if(sum(names(table(orig)) == j) != 0) {
      cucina[c, j] <- table(orig)[
        names(table(orig)) == j]/
        table(ricette$origine)[
          names(table(ricette$origine)) == j]
    } else {cucina[c, j] <- 0}
  }
}

ingr.usati.assieme <- ingredienti[ingredienti$V3 > 0 &
                                    ingredienti$V1 !=
                                    ingredienti$V2,]
```

```
cucina.coppie <- as.data.frame(
  matrix(NA, nrow = nrow(ingr.usati.assieme), ncol = 13))
colnames(cucina.coppie) <- c("ingrediente1",
                             "ingrediente2",
                             unique(ricette$origine))
cucina.coppie[,1:2] <- c(as.character(
  ingr.usati.assieme[,1]), as.character(
    ingr.usati.assieme[,2]))

for(i in 1:nrow(ingr.usati.assieme)) {
  cat(i, "\n")
  coppia <- cucina.coppie[i, 1:2]
  ricette1 <- apply(ricette, 1,
                    function(x) sum(coppia %in% x))
  ricette2 <- ricette[which(ricette1 == 2),]
  orig <- ricette2$origine
  for(j in regioni) {
    if(sum(names(table(orig)) == j) != 0) {
      cucina.coppie[i, j] <- table(orig)[
        names(table(orig)) == j]/
        table(ricette$origine)[
          names(table(ricette$origine)) == j]
    } else {cucina.coppie[i, j] <- 0}
  }
}

# numero medio composti organici per ricetta
z <- rep(0, nrow(ricette.ingr))
for(i in 1:nrow(ricette.ingr)){
```

```
  for(j in 1:ncol(ricette.ingr)) {
    if(ricette.ingr[i,j] != "") z[i] <- z[i]+1
  }
}

comp.org.ric <- as.data.frame(matrix(
  NA, nrow = nrow(ricette), ncol = 2))
comp.org.ric[, 1] <- ricette$origine

for(k in 1:nrow(ricette)) {
  mtx <- mtx.sap[as.factor(
    colnames(mtx.sap)) %in% ricette.ingr[k,],
                colnames(mtx.sap) %in% ricette.ingr[k,]]
  comp.org.ric[k, 2] <- (2/(z[k]*(z[k]-1)))*(sum(mtx)/2)
}

levels(comp.org.ric$V1) <- c("Africa",
"Asia orientale", "Europa orientale",
"America Latina", "Medio Oriente", "Nord America",
"Europa settentrionale", "Asia meridionale",
"Sud-est asiatico", "Europa meridionale",
"Europa occidentale")

comp.org.ric <- comp.org.ric %>%
                mutate(V1 = fct_relevel(comp.org.ric$V1,
                "Africa","Europa settentrionale",
                "Europa occidentale","Europa orientale",
                "Europa meridionale","Medio Oriente",
                "Sud-est asiatico","Asia meridionale",
```

```
                "Asia orientale","America Latina",
                "Nord America"))

ggplot(comp.org.ric, aes(x = V2, y = V1, fill = V1)) +
  stat_density_ridges(
                      quantile_lines = TRUE,
                      quantile_fun = median,
                      vline_color = "#d54200",
                      alpha = 1,
                      vline_size = 1

                      # quantiles = 0.5
                      ) +
  scale_fill_manual(values = c("#e5f4fa",
  "#b2def2", "#7fc9ea", "#4cb3e1",
  "#199dd9","#0084bf",
  "#006695", "#00496a","#003854",
  "#002c3f", "#000000"),
  aesthetics = "fill")+
  theme_classic() +
  theme(axis.text.x = element_text(size=20),
        axis.text.y = element_text(size=20),
        legend.position = "none") +
  scale_x_continuous(name = NULL,
    breaks=seq(0,30,5), limits = c(-1,31)) +
  scale_y_discrete(name = NULL,
    guide = guide_axis(position = "left"))
```

```
# caricato il file contenente le categorie
colnames(categorie) <- c("ingrediente", "categoria")


levels(categorie$categoria) <- c("Alcolici",
"Prodotti animali","Cereali","Latticini",
"Pesce/frutti di mare","Fiori","Frutta",
"Erbe","Carni","Noci/semi/legumi","Piante",
"Derivati vegetali","Spezie","Verdura")

categorie <- categorie %>% mutate(categoria =
                    fct_relevel(categoria,
                    "Noci/semi/legumi",
                    "Alcolici","Fiori",
                    "Frutta","Prodotti animali",
                    "Carni","Spezie",
                    "Cereali","Verdura",
                    "Derivati vegetali",
                    "Latticini",
                    "Pesce/frutti di mare",
                    "Piante","Erbe"))
```

Nella seguito verranno riportati i codici solo per le analisi svolte sui dati provenienti dalle ricette dell'Asia orientale, in quanto analoghe a quelle svolte sul dataset del Nord America.

```
# estrazione del numero di ingredienti nelle ricette
num.ingr.as <- num.ingr.reg %>% filter(
                origine == "EastAsian")
num.ingr.as <- num.ingr.as %>% select(- origine)
```

```
# dati sui composti organici
ingr.unici.as <- ricette.as %>% as.matrix(.) %>%
                    as.vector(.) %>% unique(.)
ingr.unici.as <- ingr.unici.as[- which(
                    ingr.unici.as == "")]

sapori.as <- sapori[sapori$V1 %in% ingr.unici.as &
                sapori$V2 %in% ingr.unici.as,]

# utilizzo degli ingredienti
c <- 0
saporinew <- matrix(NA, nrow = length(ingr.unici.as)*
                (length(ingr.unici.as) + 1)/2, ncol = 3)

for(i in ingr.unici.as) {
  ingr1 <- ricette.as %>% filter_all(any_vars(. %in% i))
  cat("ingrediente", c <- c+1, ":", i, "\n")
  b <- 1
  for(j in ingr.unici.as) {
    if(b <= c) {
      saporinew[((c*(c+1))/2)+b-c, 1] <- i
      saporinew[((c*(c+1))/2)+b-c, 2] <- j
      saporinew[((c*(c+1))/2)+b-c, 3] <- length(
                        which(as.vector(t(ingr1)) == j))
      b <- b+1
    }
  }
}
```

```
ingredienti.as <- as.data.frame(saporinew)
ingredienti.as$V3 <- as.numeric(ingredienti.as[, 3])

# numero ricette - diadica
mtx.ingr.as <- as.matrix(
                    get.adjacency(graph.data.frame(
                        ingredienti.as, directed = F),
                    attr = "V3"))
sum(names(diag(mtx.ingr.as))==(ingr.unici.as))

# numero ricette - nodo
utilizzo.as <- as.data.frame(ingr.unici.as)
utilizzo.as$uso <- diag(mtx.ingr.as)
diag(mtx.ingr.as) <- 0

# numero composti organici nelle ricette -
# diadica
mtx.sap <- as.matrix(
            get.adjacency(graph.data.frame(
                sapori.as, directed = F),
            attr = "V3"))
mtx.sap.as <- mtx.sap[colnames(mtx.sap) %in%
                        ingr.unici.as,
                    colnames(mtx.sap) %in%
                        ingr.unici.as]

# numero medio composti organici per ricetta
```

```
media.comp.org.as <- num_m_comp_orig %>% filter(
                .$V1 == "EastAsian") %>% select(- V1)


# n medio di composti organici nelle ricette - nodo
m.org.ric.nodo.as <- as.data.frame(matrix(NA,
            nrow = length(ingr.unici.as), ncol = 2))
colnames(m.org.ric.nodo.as) <- c("ingrediente",
                                 "composti medi")
m.org.ric.nodo.as$ingrediente <- ingr.unici.as

b <- 0
for(i in ingr.unici.as) {
  cat(b <- b+1, i, "\n")
  ricette1 <- apply(ricette.as, 1, function(x)
                                    sum(i %in% x))
  m.org.ric.nodo.as[
        m.org.ric.nodo.as$ingrediente==i,2] <-
        mean(media.comp.org.as[which(ricette1 == 1),])
}

# N MEDIO NUM MEDIO DI COMPOSTI ORGANICI - DIADICA
ingr.usati.assieme.as <- ingredienti.as[
            ingredienti.as$V3 > 0 &
            ingredienti.as$V1 != ingredienti.as$V2,]

m.org.ric.diad.as <- as.data.frame(matrix(NA,
    nrow = nrow(ingr.usati.assieme.as), ncol = 3))
colnames(m.org.ric.diad.as) <- c("ingrediente1",
```

```
                                    "ingrediente2",
                                    "composti medi")
m.org.ric.diad.as[, 1:2] <- ingr.usati.assieme.as[, 1:2]

for(i in 1:nrow(ingr.usati.assieme.as)) {
  cat(i, "\n")
  coppia <- ingr.usati.assieme.as[i, 1:2]
  ricette1 <- apply(ricette.as, 1, function(x)
                                        sum(coppia %in% x))
  m.org.ric.diad.as[i, 3]<- mean(
            media.comp.org.as[which(ricette1 == 2),])
}

# numero medio di ingredienti nelle ricette
# di nodo
m.ingr.ric.nodo.as <- as.data.frame(matrix(NA,
            nrow = length(ingr.unici.as), ncol = 2))
colnames(m.ingr.ric.nodo.as) <- c("ingrediente",
                                    "numerosità media")
m.ingr.ric.nodo.as$ingrediente <- ingr.unici.as

b <- 0
for(i in ingr.unici.as) {
  cat(b <- b+1, i, "\n")
  ricette1 <- apply(ricette.as, 1, function(x)
                                          sum(i %in% x))
  m.ingr.ric.nodo.as[m.ingr.ric.nodo.as$ingrediente==i,
        2]<- mean(num.ingr.as[which(ricette1 == 1),])
}
```

```
for(i in 1:nrow(ingr.usati.assieme.as)) {
  cat(i, "\n")
  coppia <- ingr.usati.assieme.as[i, 1:2]
  ricette1 <- apply(ricette.as, 1, function(x)
                              sum(coppia %in% x))
  m.ingr.ric.diad.as[i, 3]<- mean(num.ingr.as[
                           which(ricette1 == 2),])
}

# categoria alimentare
categorie.as <- categ[categ$ingredient.name %in%
                                  ingr.unici.as, 2:3]
colnames(categorie.as) <- c("ingrediente", "categoria")

# grafico categorie per gli ingredienti disponibili
num.categ.as <- as.data.frame(
            prop.table(table(categorie.as$categoria)))
num.categ.as[,3] <- as.data.frame(
                     table(categorie.as$categoria))[, 2]
colnames(num.categ.as) <- c("categoria",
                            "frequenza_relat",
                            "frequenza")

ggplot(data=num.categ.as, aes(x = categoria,
                y = frequenza, fill = categoria)) +
  geom_bar(stat="identity", color = "black") +
  coord_flip() +
  scale_fill_manual(values=c(
```

```
    "#ab5236","#ffa300","#ff004d","#7e2553",
    "#ff77a8","#ffccaa","#fff1e8","#f9ef9f",
    "#ffec27","#d6f264", "#a7f070", "#00e436",
    "#008751","#29adff")) +
  theme_classic() +
  theme(axis.text.x = element_text(size=20),
        axis.text.y = element_text(size=20),
        legend.position = "none") +
  scale_y_continuous(name = NULL, breaks=seq(0,54,8)) +
  scale_x_discrete(name = NULL)

# autenticità ingredienti
cucina.as <- cucina[cucina$ingrediente %in%
                                ingr.unici.as,]

# autenticità - nodo
autent.nodo.as <- as.data.frame(cucina.as$ingrediente)
colnames(autent.nodo.as) <- "ingrediente"
autent.nodo.as$autenticita <- cucina.as$EastAsian -
                        rowMeans(cucina.as %>% select(
                        - c(ingrediente, EastAsian)))

# autenticità - diadica
cucina.coppie.as <- as.data.frame(matrix(NA,
                    nrow = nrow(cucina.coppie.tot),
                    ncol = 13))

for(i in 1:nrow(cucina.coppie.tot)) {
  cat(i, "\n")
```

```
  coppia <- cucina.coppie.tot[i, 1:2]
  selez <- apply(ingr.usati.assieme.as, 1, function(x)
                                      sum(coppia %in% x))
  if(sum(selez == 2) > 0) {
    cucina.coppie.as[i,]<- cucina.coppie.tot[which(
                                          selez == 2),]
  }
}

cucina.coppie.as <- na.omit(cucina.coppie.as)
autent.diad.as <- as.data.frame(matrix(NA, nrow =
                        nrow(ingr.usati.assieme.as),
                        ncol = 3))
colnames(autent.diad.as) <- c("ingrediente1",
                            "ingrediente2",
                            "autenticita")
autent.diad.as[, 1:2] <- ingr.usati.assieme.as[, 1:2]

autent.diad.as$autenticita <- cucina.coppie.as$EastAsian -
               rowMeans(cucina.coppie.as %>% select(
            - c(ingrediente1, ingrediente2, EastAsian)))

# informazione mutua puntuale
imp.as <- as.data.frame(matrix(nrow =
                        nrow(ingr.usati.assieme.as),
                        ncol = 3))
imp.as[,1:2] <- ingr.usati.assieme.as[,1:2]
colnames(imp.as) <- c("ingrediente1",
                    "ingrediente2",
```

```
                    "imp")

mtx.p.ingr.as <- mtx.ingr.as/nrow(ricette.as)
p.utilizzo.as <- utilizzo.as
p.utilizzo.as$uso <- p.utilizzo.as$uso/nrow(ricette.as)

for(i in 1:nrow(ingr.usati.assieme.as)){
  cat(i, "\n")
  coppia <- as.character(ingr.usati.assieme.as[i,1:2])
  imp.as$imp[i] <- log(mtx.p.ingr.as[coppia[1],
        coppia[2]]/(p.utilizzo.as[
        p.utilizzo.as$ingr.unici.as==coppia[1],2]*
        p.utilizzo.as[
        p.utilizzo.as$ingr.unici.as==coppia[2],2]))
}
```

Analisi esplorative.

```
# ricette
ricette.tot <- ricette.
ricette.as.orig <- ricette.tot %>% filter(
                            origine == "EastAsian")
ricette.as <- ricette.as.orig %>% select(- origine)

# media numero medio di composti organici - nodo
b <- 0
for(i in ingr.unici.as) {
  cat(b <- b+1, i, "\n")
  ricette1 <- apply(ricette.as, 1, function(x)
                                        sum(i %in% x))
```

```
  m.org.ric.nodo.as[m.org.ric.nodo.as$ingrediente==i,
    2]<-mean(media.comp.org.as[which(ricette1 == 1),])
}

# ingredienti più usati e più autentici
utilizzo.as %>% slice_max(order_by = .$uso, n = 10) %>%
                        cbind(.$uso/nrow(ricette.as))
autent.nodo.as[autent.nodo.as$ingrediente %in%
        utilizzo.as$ingrediente[
        utilizzo.as$uso/nrow(ricette.as) > 0.01],] %>%
        slice_max(., order_by = .$autenticita, n = 10)

# ingredienti autentici
aut.as <- autent.nodo.as[autent.nodo.as$ingrediente %in%
utilizzo.as$ingrediente[utilizzo.as$uso/
nrow(ricette.as) > 0.01],] %>%
filter(., .$autenticita > 0)

# diadi autentiche
diad.pres.as <- ingredienti.as[ingredienti.as$uso/
                            nrow(ricette.as) > 0.01 &
                        ingredienti.as$ingrediente1 !=
                        ingredienti.as$ingrediente2,]
diad.aut.as <- left_join(diad.pres.as[,1:2],
                autent.diad.as,
                by = c("ingrediente1", "ingrediente2"))
diad.aut.as %>% slice_max(order_by = .$autenticita,
                                                n = 10)
aut2.as <- diad.aut.as[diad.aut.as$autenticita > 0, ]
```

```
nrow(aut2.as)
nrow(ingr.usati.assieme.as)
aut2.as[,1:2] %>% as.matrix(.) %>% as.vector(.) %>%
                                    unique(.) %>% length(.)

# categorie
categ.aut.as <- left_join(aut.as, categorie.as,
                                by = "ingrediente")

# pie chart
library(ggplot2)
colnames(utilizzo.as) <- c("ingrediente", "uso")
pie.as <- inner_join(utilizzo.as, categorie.as,
        by = "ingrediente") %>% select(- "ingrediente")
pie.as <- aggregate(uso ~ categoria, pie.as, FUN = sum)
pie.as %>% ggplot(aes(x = "", y = uso,
    fill = categoria)) +
  geom_bar(stat="identity", color = "black") +
  coord_polar("y", start=20) +
  scale_fill_manual(values=c(
    "#cc9966","#ffccff","#cc66ff",
    "#ff33ff","#ff6699","#cc0033",
    "#ff9933","#ffcc00","#00ff33",
    "#99ff66","#99ffff","#66ccff",
    "#669966","#999900")) +
  theme_minimal() +
  ggtitle("Asia orientale")+
  theme(axis.text.x=element_blank(),
        # legend.position = "none",
```

```
        plot.title = element_text(size=40,
        hjust = 0.5, vjust = -2)) +
  scale_x_discrete(name = NULL) +
  scale_y_discrete(name = NULL)

comp.ric.as <- left_join(aut.as, m.org.ric.nodo.as,
                                    by = "ingrediente")
comp.ric2.as <- left_join(aut2.as, m.org.ric.diad.as,
            by = c("ingrediente1","ingrediente2"))

# composti organici
# matrice di adiacenza e rete per i composti organici
mtx.c.org <- sapori
mtx.c.org <- as.matrix.data.frame(mtx.c.org,
                                    rownames.force = T)

uso.autent.as <- left_join(autent.nodo.as,
                                  m.org.ric.nodo.as)
uso.autent.as <- left_join(uso.autent.as, categorie.as)
uso.autent.as <- left_join(uso.autent.as, utilizzo.as)
uso.autent.as <- uso.autent.as %>% filter(
                        .$uso/nrow(ricette.as) > 0.01)
uso.autent.as$ingrediente <- c("mandorla",
"orzo","fagioli","manzo","brodo di manzo",
"peperone","fagioli neri","pepe nero",
"brassica","pane","broccolo","grano saraceno",
"burro","cavolo","melassa di canna","carota",
"peperoncino","sedano","olio di sedano","pollo",
"brodo di pollo","cavolo cinese","coriandolo",
```

```
"cannella","vongola","coriandolo","granturco",
"granchio","cetriolo","uovo","enokidake","pesce",
"aglio","zenzero","prosciutto","miele","katsuobushi",
"kelp","limone","succo di limone","lattuga","sgombro",
"matsutake","carne","latte","funghi","cozza","senape",
"nira","frutta secca","polpo","olio d'oliva","cipolla",
"ostrica","prezzemolo","piselli","arachide",
"olio di arachidi","pera","pepe","maiale","patata",
"zucca","ravanello","fagioli rossi","vino rosso",
"riso","semi di sesamo tostati","sake","cipollotto",
"alghe marine","olio di sesamo","semi di sesamo","scalogno",
"sherry","shiitake","gambero","salsa di soia","soia",
"calamaro","amido","patata dolce","pomodoro","tonno",
"vegetali","olio vegetale","aceto","wasabi","grano",
"vino bianco","vino","lievito")

uso.autent.as %>% ggplot(aes(x = autenticita,
    y = `composti medi`, fill = categoria,
    label=ingrediente)) +
  geom_point(shape=23) +
  geom_label(
    # aes(size=uso)
    )+
  # geom_text(aes(size=uso))
  # coord_polar("y", start=20) +
  scale_fill_manual(values=c(
    "#cc9966","#ffccff","#cc66ff",
    "#ff33ff","#ff6699","#cc0033",
    "#ff9933","#ffcc00","#00ff33",
```

```
    "#99ff66","#99ffff","#66ccff",
    "#669966","#999900")) +
  theme_minimal() +
  ggtitle("Asia orientale")+
  theme(
    # axis.text.x=element_blank(),
        legend.position = "none",
        plot.title = element_text(size=20,
        hjust = 0.5, vjust = 0)
        ) +
  scale_y_continuous(name = "Media dei composti organici
                delle ricette", breaks=seq(0,60,10)) +
  scale_x_continuous(name = "Autenticità",
                             breaks=seq(-0.4,0.7,0.2))

mtx.c.org.as <-as.matrix(get.adjacency(graph.data.frame(
  sapori.as, directed = F), attr = "comp_org"))
rete.c.org.as <- graph.adjacency(mtx.c.org.as,
                             mode = "undirected",
                             weighted = T, diag = F)

edge_density(rete.c.org.as)
gorder(rete.c.org.as)
gsize(rete.c.org.as)
degree.as <- degree(rete.c.org.as)
strength.as <- strength(rete.c.org.as, loops = F)
ClustBCG(mtx.c.org.as)
ClustF(mtx.c.org.as)
```

```
# ingredienti autentici
mtx.c.org.aut.as <- mtx.c.org.as[
  colnames(mtx.c.org.as) %in% aut.as$ingrediente,
  colnames(mtx.c.org.as) %in% aut.as$ingrediente]

rete.c.org.aut.as <- graph.adjacency(mtx.c.org.aut.as,
        mode = "undirected", weighted = T, diag = F)

# indici descrittivi
gorder(rete.c.org.aut.as)
gsize(rete.c.org.aut.as)
diameter(rete.c.org.aut.as, directed = F)
edge_density(rete.c.org.aut.as)
strength.aut.as <- strength(rete.c.org.aut.as, loops = F)

# creazione dataset per il grafo
nodi_grafo <- left_join(aut.as, categ %>%
                select(- X..id), by = "ingrediente")
nodi_grafo <- mutate(nodi_grafo, id = row_number())
colnames(nodi_grafo) <- c("label",
                          "weight",
                          "class",
                          "id")

connect <- sapori.as %>% filter(.$ingrediente1 %in%
        nodi_grafo$label) %>%
         filter(.$ingrediente2 %in% nodi_grafo$label)
connect <- left_join(connect, nodi_grafo %>% select(
    -c(weight, class)), by = c("ingrediente1"="label"))
```

```
connect <- left_join(connect, nodi_grafo %>% select(
    -c(weight, class)), by = c("ingrediente2"="label"))
archi_grafo <- connect %>% select(-c(ingrediente1,
    ingrediente2)) %>% mutate(., id = row_number())
colnames(archi_grafo) <- c("weight","source",
                           "target","id")
nomi.ita <- as.factor(c("orzo",
"manzo","brodo di manzo","fagioli neri","pepe nero",
"brassica","broccoli","grano saraceno","cavolo",
"carota","pepe di caienna","olio di sedano",
"cavolo cinese","vongole","granchio","cetriolo",
"enoki","pesce","aglio","zenzero","katsuobushi",
"alghe kombu","lattuga","sgombro","matsutake",
"funghi","cozze","nira","noce","polpo","ostriche",
"arachidi","olio di arachidi","pera","maiale",
"zucca","ravanello","fagioli rossi","riso",
"semi di sesamo tostati","sake","cipollotto",
"alghe","olio di sesamo","semi di sesamo",
"sherry","shiitake","gamberetti","salsa di soia",
"soia","calamari","amido","patata dolce",
"tonno","verdura","olio vegetale","aceto",
"wasabi","vino"))
nodi_grafo$label <- nomi.ita

# rete dell'informazione mutua puntuale
mtx.imp.as <- as.matrix(get.adjacency(graph.data.frame(
                imp.as, directed = F), attr = "imp"))
rete.imp.as <- graph.adjacency(mtx.imp.as,
            mode = "undirected", weighted = T, diag = F)
```

```
edge_density(rete.imp.as)
degree.as <- degree(rete.imp.as)
strength.as <- strength(rete.imp.as, loops = F)
ClustBCG(mtx.imp.as)
ClustF(mtx.imp.as)

# rete imp positivi
imp.asp <- imp.as[imp.as$imp>0,]
mtx.imp.asp <- as.matrix(get.adjacency(graph.data.frame(
  imp.asp, directed = F), attr = "imp"))
rete.imp.asp <- graph.adjacency(mtx.imp.asp,
        mode = "undirected", weighted = T, diag = F)
edge_density(rete.imp.asp)
strength.asp <- strength(rete.imp.asp, loops = F)
ClustBCG(mtx.imp.asp)
ClustF(mtx.imp.asp)

# rete imp negativi
imp.asn <- imp.as[imp.as$imp<0,]
mtx.imp.asn <- as.matrix(get.adjacency(graph.data.frame(
  imp.asn, directed = F), attr = "imp"))
rete.imp.asn <- graph.adjacency(mtx.imp.asn,
        mode = "undirected", weighted = T, diag = F)
edge_density(rete.imp.asn)
strength.asn <- strength(rete.imp.asn, loops = F)
ClustBCG(mtx.imp.asn)
ClustF(mtx.imp.asn)
```

Stima e valutazione dei modelli.

```
colnames(mtx.sap.as)[ingr.unici.as %in%
                     colnames(mtx.sap.as)==F]
mtx.sap <- as.matrix(get.adjacency(graph.data.frame(
        sapori.as, directed = F), attr = "comp_org"))
mtx.sap <- cbind(mtx.sap, 0, 0)
mtx.sap <- rbind(mtx.sap, 0, 0)
colnames(mtx.sap)[241:242] <- colnames(
                     mtx.sap.as)[ingr.unici.as %in%
                     colnames(mtx.sap.as)==F]
rownames(mtx.sap)[241:242] <- rownames(mtx.sap.as)[
        ingr.unici.as %in% rownames(mtx.sap.as)==F]
mtx.sap <- mtx.sap[order(
        colnames(mtx.sap)),order(colnames(mtx.sap))]

mtx.aut <- as.matrix(get.adjacency(graph.data.frame(
                     autent.diad.as, directed = F),
                     attr = "autenticita"))
for(i in autent.nodo.as$ingrediente) {
  mtx.aut[colnames(mtx.aut)==i,
colnames(mtx.aut)==i] <- autent.nodo.as$autenticita[
            autent.nodo.as$ingrediente==i]
}
mtx.aut <- mtx.aut[order(colnames(mtx.aut)),
                     order(colnames(mtx.aut))]

mtx.imp <- as.matrix(get.adjacency(graph.data.frame(
            imp.as, directed = F), attr = "imp"))
mtx.imp <- mtx.imp[order(colnames(mtx.imp)),
                order(colnames(mtx.imp))]
```

```
mtx.dim <- as.matrix(get.adjacency(graph.data.frame(
        m.ingr.ric.diad.as, directed = F),
        attr = "numerosità.media"))
for(i in m.ingr.ric.nodo.as$ingrediente) {
  mtx.dim[colnames(mtx.dim)==i,
  colnames(mtx.dim)==i] <- m.ingr.ric.nodo.as$`numerosità media`[
  m.ingr.ric.nodo.as$ingrediente==i]
}
mtx.dim <- mtx.dim[order(colnames(mtx.dim)),
            order(colnames(mtx.dim))]

mtx.org.ric <- as.matrix(get.adjacency(graph.data.frame(
            m.org.ric.diad.as, directed = F),
            attr = "composti.medi"))
for(i in m.org.ric.nodo.as$ingrediente) {
  mtx.org.ric[colnames(mtx.org.ric)==i,
  colnames(mtx.org.ric)==i] <- m.org.ric.nodo.as$`composti medi`[
  m.org.ric.nodo.as$ingrediente==i]
}
mtx.org.ric <- mtx.org.ric[order(colnames(mtx.org.ric)),
            order(colnames(mtx.org.ric))]

X <- array(dim = c(nrow(mtx.imp),ncol(mtx.imp),5))
X[,,1] <- mtx.imp
X[,,2] <- mtx.sap
X[,,3] <- mtx.org.ric
X[,,4] <- mtx.aut
X[,,5] <- mtx.dim
```

```
X2 <- array(dim = c(nrow(mtx.imp),4))
X2[,1] <- categorie.as$categoria[order(
            categorie.as$ingrediente)]
X2[,4] <- m.ingr.ric.nodo.as$`numerosità media`[order(
                m.ingr.ric.nodo.as$ingrediente)]
X2[,2] <- m.org.ric.nodo.as$`composti medi`[order(
            m.org.ric.nodo.as$ingrediente)]
X2[,3] <- autent.nodo.as$autenticita[order(
            autent.nodo.as$ingrediente)]

rete <- list(diadiche =array(
        dim = c(nrow(mtx.imp),ncol(mtx.imp),5)),
        nodo = array(dim = c(nrow(mtx.imp), 4)))
rete$diadiche <- X
rete$nodo <- X2
dimnames(rete$diadiche)[[1]] <- rownames(mtx.imp)
dimnames(rete$diadiche)[[2]] <- rownames(mtx.imp)
dimnames(rete$diadiche)[[3]] <- c("imp",
                                  "sapori",
                                  "c.org.ric",
                                  "autent",
                                  "dim.ric")
dimnames(rete$nodo)[[1]] <- rownames(mtx.imp)
dimnames(rete$nodo)[[2]] <- c("categ",
                              "c.org.ric",
                              "autent",
                              "dim.ric")

gofstats(rete$diadiche[,,1])
```

```
#modello SRRM
fit_SRRM.as <- ame(rete$diadiche[,,1],
                   Xd = rete$diadiche[,,2:5],
                   Xr = rete$nodo,
                   family = "nrm",
                   symmetric = T, plot=F)
summary(fit_SRRM.as)
plot(fit_SRRM.as)

for(i in c(1:15)){
  fit_AME <- ame(Y=rete$diadiche[,,1],
                 Xd=rete$diadiche[,,2:5],
                  Xr = rete$nodo,
                  family="nrm", symmetric=T, R=i,
                  nvar = T,
                  plot=F, print=T)
  save(fit_AME, file=paste0("~/ame.as.", i, ".RData"))
}

srrm.as <- fit_SRRM.as
plot(srrm.as)
summary(srrm.as)

ame.as <- fit_AME
plot(ame.as)
summary(ame.as)
apply(ame.as$GOF, 2, function(x) mean(x))
ame.as$L
```