Incidenti sul lavoro negli USA: studio del rischio nei vari settori

Fabio Marigo[1], Federico Melograna[1], Silvia Santamaria[2], Michela Sessi[2], Gianmarco Stucchi[2]

[1] *CdLM CLAMSES, Università degli studi di Milano Bicocca*
[2] *CdLM Data Science, Università degli studi di Milano Bicocca*

Come migliorare la sicurezza e la sanità dell'ambiente lavorativo negli USA? Questa è la domanda che si pone ogni anno la Occupational Safety and Health Administration (OSHA) che dal 1 Gennaio 2015 ha richiesto obbligatoriamente un rapporto ai datori di lavoro per ogni grave infortunio di un loro dipendente. Questa recente fonte di dati può portare alla realizzazione di una strategia per l'anno a seguire: valutare e comprendere gli incidenti sono elementi fondamentali per la pianificazione e l'elaborazione di politiche della sicurezza e della sanità a livello geografico e ancor più a livello settoriale. Sulla base di tale affermazione si sviluppa questo studio: attraverso una Cluster Analysis si cerca di evidenziare eventuali gruppi di settori con elevato rischio di incidenti sul luogo di lavoro in base a caratteristiche dell'evento e della causa scatenante.

1 Introduzione

Dal 1970 negli USA la Occupational Safety and Health Administration (OSHA) [1] si impegna per creare un ambiente di lavoro sicuro e sano. Dal 2015 richiede che i datori di lavoro segnalino i gravi infortuni dei loro dipendenti, definiti come un'amputazione, un ricovero in ospedale o la perdita di un occhio. I dati disponibili sono aggiornati a Febbraio 2017 e superano i 20 mila casi di incidenti. Nei rapporti vengono descritti gli incidenti, fornendo sia informazioni temporali e geografiche sia le caratteristiche - tramite codifica OIICS[1] - degli infortuni. L'analisi di tali rapporti e la comprensione degli incidenti sono elementi indispensabili per la pianificazione e l'elaborazione di politiche del territorio per la ridistribuzione di fondi e per l'organizzazione di campagne di prevenzione secondo un piano coerente con le sue esigenze. Tale analisi deve avvalersi di valutazioni tanto geografiche quanto settoriali: non solo la differenziazione del luogo ma ancor più la variazione della tipologia di stabilimento possono influire sugli infortuni.

Questo lavoro nasce con lo scopo di indagare il potenziale rischio settoriale delle aziende e industrie degli USA a partire dal 2015, definendo il rischio come rapporto tra il numero di incidenti e il numero di dipendenti che lavorano in un determinato settore. Per elaborare una tale classificazione si è usato il North American Industry Classification System (NAICS) [3] - che identifica univocamente ogni azienda in base al campo in cui compete - e si è effettuata un'aggregazione dei casi. Si vuole quindi effettuare e validare una Cluster Analysis che metta in luce eventuali gruppi di settori (e quindi aziende) con elevato rischio di incidenti sul posto di lavoro in base ad alcune informazioni e codifiche sugli infortuni.

[1] Occupational Injury and Illness Classification System (OIICS) [2]

2 Descrizione dataset

L'analisi nasce con l'idea di confrontare il numero e le caratteristiche degli incidenti con l'appartenenza ai corrispondenti settori aziendali. Per questo motivo i dati utilizzati per le analisi fanno riferimento a due diverse fonti:

- la prima contiene i rapporti dei datori di lavoro sugli infortuni[2];
- la seconda riporta la classificazione e alcune informazioni sui settori aziendali[3].

2.1 Severely Injured Workers

I primi dati sono reperibili nel progetto *Severely Injured Workers* reso fruibile dalla piattaforma Kaggle[4] contenente due dataset. Il primo, `severeinjury.csv`, raccoglie le segnalazioni dei datori di lavoro dal 1 Gennaio 2015 al 28 Febbraio 2017, per un totale di oltre 20mila casi di incidenti gravi ma non mortali. Per ogni rapporto si ottengono informazioni quali data, diversi dettagli sul luogo geografico, nome e codice NAICS dell'azienda, eventuale ricovero in ospedale e/o amputazioni, descrizione dell'infortunio e quattro codici relativi alla classificazione degli incidenti OIICS mirati a classificarne le tipologie per evento, causa e parte del corpo lesa. In particolare per valutare i codici si usa il secondo dataset (`oiics_201_code_list.xlsx`) che riporta un dizionario di questi quattro attributi resi fruibili a quattro livelli di dettaglio.

[2] Per maggiori informazioni si veda l'appendice A.
[3] Per maggiori informazioni si veda l'appendice B.

2.2 United States Census

Il *North American Industry Classification System* (NAICS)[3] è una codifica usata da imprese e governo americani allo scopo di identificare il tipo di attività svolta; la numerazione è composta al massimo da sei cifre, che caratterizzano l'impresa nel modo più dettagliato possibile. Le prime tre cifre indicano il sottosettore al quale lo stabilimento appartiene. Le informazioni sono prese dal sito *United States Census*, il quale permette l'accesso e il download di numerosi dati, grazie alla piattaforma accessibile liberamente *American FactFinder*. Tramite questo *tool* si è creata una tabella che incrocia le informazioni relative ai sottosettori con alcune statistiche industriali risalenti al 2015, quali numero di stabilimenti, numero di impiegati e paghe totali annuali. Il dataset ottenuto è composto dunque da 86 osservazioni.

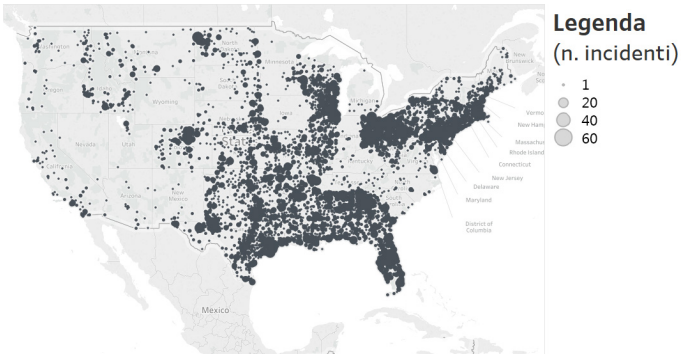

Figura 1: *Distribuzione geografica degli oltre 20mila incidenti*

3 Ristrutturazione dataset

Per lo svolgimento del lavoro si è utilizzato la piattaforma Knime[5] con l'appoggio del linguaggio di programmazione R[6]: i due programmi possono dialogare tra loro tramite l'utilizzo del nodo RSnippet[4]. Si è creato un unico dataset dall'unione delle due fonti, Severely Injured Workers e US Census. La scelta è stata dettata da molteplici fattori:

- la gestione di un numero elevato di casi porta a problemi computazionali che richiedevano prestazione troppo elevate;
- l'utilizzo di due dataset distinti porta a problemi pratici in quanto la piattaforma Knime talvolta non permette l'inserimento di due dataset distinti;
- il dataset *SeverelyInjuredWorkers*, composto solo da variabili nominali, permette l'implementazione di un numero ristretto di modelli;
- le variabili relative alle diverse caratteristiche degli infortuni (*Nature, PartofBody, Event, Source*), costituite da codificazione OIICS, presentano numerose modalità. Questo aspetto in fase di binarizzazione dei dati implica un aumento esponenziale della numerosità delle variabili.

[4]per le analisi si sono utilizzati i pacchetti *cluster*, *class*, *cld*, *plyr*, *ggplot2*, *ggdendro*, *clValid*

3.1 Definizione e selezione delle variabili

Per la creazione del nuovo dataset (`naics.csv`) si sono considerate 3 cifre del codice NAICS, mantenendo quindi il livello di dettaglio scelto nella formazione del dataset settoriale. I dati dei sottosettori relativi al 2015 permettono di uniformare le analisi in tutto l'arco temporale considerato, imponendo un punto di riferimento comune. Successivamente si sono calcolate le frequenze assolute degli infortuni per ogni sottosettore, ottenendo così la variabile *Casi* (C). I sottosettori che presentano un valore di *Casi* estremamente basso (minore di 5) vengono esclusi dal nuovo dataset, che quindi è formato in totale da 79 sottosettori.

Analogamente a quanto fatto per la variabile *Casi*, sono state calcolate le occorrenze delle modalità di *Ricovero* (RCV) e *Amputazione* (AMP) condizionatamente ai sottosettori. Si è esclusa la variabile *PartofBody* in quanto non si ritiene vi sia una correlazione tra la parte del corpo interessata e il rischio associato al settore. Inoltre più del 70% delle parti del corpo interessate dagli infortuni riguardano gli arti superiori ed inferiori.

Della variabile *Nature* si sono considerate solo le modalità relative ad eventi traumatici[5]. Le variabili relative alle caratteristiche degli infortuni (codifiche OIICS) sono state ridotte a due cifre per correggere l'eccessiva sparsità dei dati. Tramite una procedura di aggregazione si sono quindi ottenute cinque variabili:

- Infortuni traumatici (N_1): numero di infortuni che interessano fratture ossee, stiramenti muscolari, lesione ai legamenti
- Ferite aperte (N_2): numero di incidenti da cui sono scaturiti tagli, amputazioni, lacerazioni
- Ustioni (N_3): numero di casi in cui il soggetto ha riportato diversi gradi di bruciature
- Trauma cranico (N_4): numero di infortuni con presenza di emorragie celebrali, colpi alla testa
- Condizioni avverse (N_5): numero di casi in cui il soggetto ha riportato forti mal di testa, cali di pressione dovuti ad ambienti eccessivamente caldi o freddi, ipotermia ed eccessivo affaticamento

La variabile *Event* ha dato origine a sette nuove variabili, considerando solo la prima cifra del codice OIICS:

- Violenza (E_1): numero di infortuni dovuti a violenze fisiche causate da persone o animali
- Trasporto (E_2): numero di incidenti aerei, ferroviari e stradali.
- Esplosioni (E_3): numero di infortuni avvenuti in presenza di incendi, esplosioni di macchinari, demolizione di strutture
- Cadute (E_4): numero di infortuni dovuti a scivolate, cadute da scale o strutture
- Esposizione ad energie (E_5): numero di infortuni dovuti a forti scosse elettriche o radiazioni
- Contatto (E_6): numero di infortuni dovuti a urti violenti con oggetti, equipaggiamenti o corpi contundenti
- Affaticamento (E_7): numero di incidenti dovuti ad eccessivo affaticamento dal punto di vista fisico

[5]dove la prima cifra del codice OIICS è pari ad 1

Lo stesso procedimento effettuato per la precedente variabile è stato fatto per la variabile *Source* che si riferisce alla fonte dell'incidente. Si sono ottenute 8 variabili:

- Chimica (S_1): numero di incidenti in cui sono coinvolti prodotti o sostanze chimiche
- Containers (S_2): numero di incidenti in cui sono coinvolti contenitori, apparecchi, arredi, ecc.
- Macchinari (S_3): numero di incidenti avvenuti durante l'utilizzo di un macchinario
- Materiali (S_4): numero di incidenti avvenuti durante l'utilizzo o lo spostamento di materie prime come legno, plastiche, metalli, ecc.
- Esseri viventi (S_5): numero di incidenti in cui sono coinvolte persone, piante o animali
- Strutture (S_6): numero di incidenti avvenuti in determinati luoghi come tunnel, cave, dighe ecc.
- Strumenti (S_7): numero di incidenti legati all'utilizzo di strumenti come coltelli, forbici, scalpelli, ecc.
- Veicoli (S_8): numero di incidenti avvenuti su veicoli stradali e non

Poiché la variabile *Casi* (C) è influenzata dalla numerosità dei lavoratori, una valida misura del rischio è rappresentata dal rapporto tra i casi di infortunio e la variabile *Impiegati* (I), ovvero il numero totale di dipendenti che lavora in quello specifico sottosettore:

$$Rischio\ (RSK) = \frac{C}{I}$$

Infine nel nuovo dataset è stata inserita la variabile *Paghe* (P), utilizzata in seguito per degli approfondimenti.

4 Clustering

Per classificare i settori secondo il rischio associato si procede con la Cluster Analysis con 2 differenti metodi di aggregazione: *gerarchico* e *K-medoids*. Si sottopone a tali tecniche la matrice $\mathbf{X}$ così definita[6]:

$$\mathbf{X} = (N_1, \ldots, N_5, E_1, \ldots, E_7, S_1, \ldots, S_8)$$

4.1 Metodi gerarchici, Ward

I Cluster gerarchici possono essere di due diversi tipi:

- agglomerativi: inizialmente si considera ciascuna osservazione come un cluster; in seguito, con una procedura a *step*, si raggruppano gradualmente i record fino ad arrivare ad un unico cluster
- divisivi: inizialmente le osservazioni formano un unico cluster, il quale negli *step* successivi viene suddiviso in cluster più piccoli.

Per individuare i cluster più simili da aggregare si è utilizzato il metodo agglomerativo di Ward (o della devianza minima). Con questo metodo ogni osservazione inizialmente forma un cluster, e ad ogni passo dell'algoritmo si fondono due cluster alla volta in modo da minimizzare l'incremento della devianza nei cluster. Per fare ciò, l'algoritmo necessita come input la matrice delle distanze $\mathbf{D} = \text{dist}(\mathbf{X})$, calcolata sulla base della distanza euclidea.

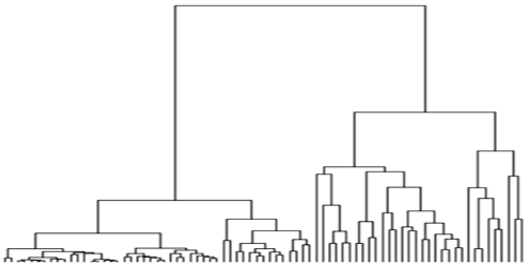

Figura 2: *Dendrogramma*

Per la scelta del numero di cluster da considerare si è osservato il dendrogramma (Figura 2) che mostra le relazioni tra cluster e sotto-cluster e l'ordine secondo il quale i cluster vengono aggregati. In particolare, sull'asse delle ordinate viene riportato il livello di distanza tra gruppi, mentre sull'asse delle ascisse vengono riportate le singole osservazioni. Ogni ramo del dendrogramma (linea verticale) corrisponde ad un cluster.

Con l'obiettivo di massimizzare la distanza tra gruppi il dendrogramma potrebbe essere tagliato così da formare due cluster; tuttavia, si nota che il cluster di destra è a sua volta scomponibile in altri due sotto-cluster ben distinti. Considerando sia questo motivo sia l'obiettivo di mantenere più informazioni possibili, il dendrogramma viene tagliato più in basso così da ottenere tre cluster.

4.2 K-Medoids

L'algoritmo K-Medoids segue una procedura iterativa: inizialmente vengono scelte casualmente K osservazioni come medoidi. Dopodiché si formano K cluster attribuendo ogni osservazione al medoide più simile secondo la distanza euclidea. Dopo aver calcolato per ciascun cluster il suo medoide (ovvero l'oggetto localizzato più centralmente) si formano nuovamente dei cluster unendo le osservazioni al medoide più simile. L'algoritmo si arresta quando converge e quindi da un'iterazione all'altra i medoidi restano invariati. L'algoritmo K-Medoids, anche se computazionalmente oneroso, è meno sensibile agli outliers; quindi, in certe situazioni (come quella in analisi) in cui è necessario non rimuovere gli outliers, può dare risultati migliori rispetto ad altri algoritmi più sensibili (come ad esempio il K-Means).

Per stabilire il numero di cluster ottimale (K), oltre a tenere in considerazione il risultato del cluster gerarchico, si calcolano il coefficiente di Silhouette e l'indice di Dunn che verranno approfonditi nel capitolo 5.

Entrambi gli indici devono essere massimizzati e quindi, come si osserva dalla figura 3, il numero di cluster ottimale è pari a 2 o 3. Così come per Ward, con l'obiettivo di utilizzare il maggior numero di informazioni disponibili, si suddivide il dataset in 3 cluster.

[6] Le variabili contenute in tale matrice sono state sottoposte ad una procedura di standardizzazione

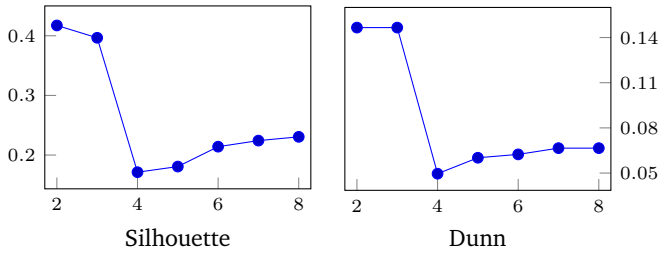

Figura 3: *Variazione del coefficiente di Silhouette e dell'indice di Dunn in relazione alla variazione del numero di cluster*

5 Validazione

Anche se la validazione dei cluster non è un campo ben definito della cluster analysis, essa è di fondamentale importanza. Infatti ciascun algoritmo che ha come fine l'individuazione dei cluster, separa le osservazioni anche nel caso in cui non vi sia alcuna struttura nei dati; pertanto, è necessario valutare la bontà dei risultati. A questo scopo si sono considerate due tipologie di indicatori: esterni (supervisionati) e interni (non supervisionati).

5.1 Indici esterni o supervisionati

Per poter mettere a confronto l'output dei due algoritmi scelti, e allo stesso tempo verificare che all'interno dei tre clusters siano raggruppati veramente sottosettori con un livello di *Rischio* simile, si dà vita alla variabile *Risk Binned*, ottenuta discretizzando *Rischio* (RSK), la quale dopo essere stata opportunamente normalizzata in $[0, 1]$ viene suddivisa in 5 intervalli di pari ampiezza. Pertanto considerando la partizione

$$RSK_B = \{\text{Basso}, \text{Medio-Basso}, \text{Medio}, \text{Medio-Alto}, \text{Alto}\}$$

e le suddivisioni nei cluster ottenute precedentemente, è possibile distinguere quattro situazioni per ciascuna coppia di osservazioni x e y:

(a) x e y appartengono allo stesso cluster e allo stessa partizione;
(b) x e y appartengono allo stesso cluster ma non alla stessa partizione;
(c) x e y appartengono alla stessa partizione ma non allo stesso cluster;
(d) x e y appartengono a cluster e partizione diversi.

Quindi sono stati applicati ai diversi algoritmi di cluster i seguenti indici:

1. **Rand**, definito come $R = \dfrac{(a+d)}{79}$
2. **Jaccard**, definito come $J = \dfrac{a}{a+b+c}$
3. **Fowlkes and Mallows**, definito come

$$FM = \sqrt{\frac{a}{a+b}\frac{a}{a+c}}$$

Indice	R	J	FM
Ward	0.904	0.802	0.890
K-Medoids	0.904	0.804	0.891

Tabella 1: *Indici esterni o supervisionati*

I risultati ottenuti sono riportati nella Tabella 1

Si osserva che gli algoritmi sono caratterizzati da indicatori con valori tanto alti quanto simili tra di loro. Poiché non è possibile scegliere l'algoritmo migliore solo sulla base di questi indici, si prosegue prendendo in considerazione gli indici interni o non-supervisionati.

5.2 Indici interni o non-supervisionati

La maggior parte degli indici interni di validazione giudica la bontà dei cluster basandosi sulla nozione di *Cohesion* (quanto sono simili le osservazioni che appartengono allo stesso cluster) e *Separation* (quanto sono distanti osservazioni che appartengono a cluster diversi).

Si sono utilizzati due diversi indici che sono una combinazione di queste due nozioni:

1. il **coefficiente di Silhouette**, definito come:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \in [-1, 1]$$

dove

- a_i rappresenta la distanza media tra la i-esima osservazione e le altre osservazioni appartenenti allo stesso cluster;
- b_i rappresenta il valore minimo della media delle distanze tra la i-esima osservazione e le osservazioni appartenenti a cluster diversi da quello a cui appartiene la i-esima osservazione;

2. il **coefficiente di Dunn**, definito in $[0, +\infty]$. Per un approfondimento circa questo indicatore si rimanda al paper [7].

I risultati ottenuti sono sintetizzati nella Tabella 2

Indice	Silhouette	Dunn
Ward	0.241	0.062
K-Medoids	0.397	0.147

Tabella 2: *Indici interni o non-supervisionati*

Valori alti per entrambi i coefficienti sono sinonimo di un buon funzionamento degli algoritmi utilizzati, in particolare il valore di Silhouette dovrebbe essere positivo. Confrontando i risultati, si rileva che il metodo di Ward manifesti prestazioni inferiori all'algoritmo K-Medoids sui dati in esame, per cui si sceglie l'output quest'ultimo per il proseguo delle valutazioni.

5.3 Test formale

Allo scopo di verificare che sussista un'effettiva struttura all'interno dei dati, si esegue un test in cui l'ipotesi nulla H_0 consiste nell'affermare che le posizioni delle osservazioni del dataset in una regione di un piano n-dimensionale siano equiprobabili (*Random Position Hypothesis*).

Il coefficiente di Silhouette relativo all'algoritmo K-Medoids precedentemente calcolato viene utilizzato per testare H_0. In particolare, applicando il metodo computazionale di Monte Carlo, si genera una distribuzione empirica sulla base di 1000 simulazioni; il quantile di questa distribuzione, che dipende dal livello di significatività scelto ($\alpha = 0.01$), viene confrontato con il valore della statistica-test (ovvero il coefficiente di Silhouette).

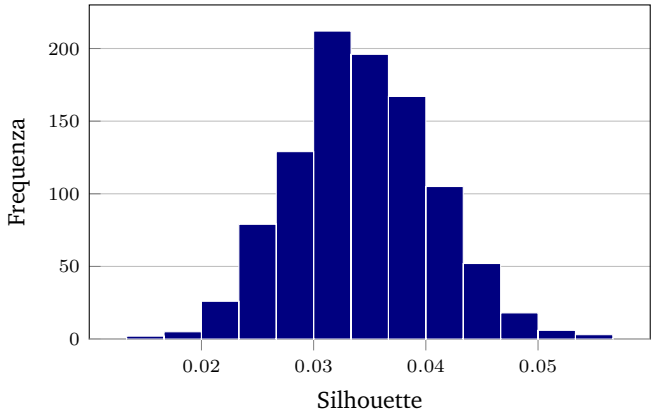

Figura 4: *Distribuzione empirica dell'indice di Silhouette ottenuta con il metodo Monte Carlo*

Poiché il quantile (pari a 0.05) è nettamente inferiore al vero valore del coefficiente di Silhouette (0.397), si rifiuta l'ipotesi di assenza di struttura e si conclude che la soluzione di clustering proposta abbia un significato.

6 Interpretazione dei risultati

In seguito all'analisi di clustering ci si trova di fronte a tre gruppi di sottosettori, ognuno dei quali con le proprie peculiarità. In questa sezione del report si vogliono indagare alcune relazioni che intercorrono tra i tre cluster e alcune delle variabili di interesse del dataset. Nello specifico ci si concentra su come varia l'incidenza del numero di infortuni, in particolare del sottoinsieme di casi più gravi ovvero quelli che portano ad un ricovero del lavoratore, oppure al mutilamento di una o più parti del corpo. Un altro punto di interesse è se sussiste o meno una remunerazione del rischio per il lavoratore.

6.1 Incidenza di infortuni, ricoveri e amputazioni nei cluster

Oltre alla già trattata variabile *Rischio* (RSK), vengono prese in considerazione le variabili *Ricovero* (RCV) e *Amputazione* (AMP), entrambe messe a rapporto con il numero di impiegati (I). In questo modo si ottengono delle misure del "rischio di ricovero" e del "rischio di amputazione":

$$RCV' = \frac{RCV}{I}; \qquad AMP' = \frac{AMP}{I}$$

Dopodiché, per una maggiore leggibilità dei risultati, le due nuove variabili create e *Rischio* vengono normalizzate tramite la seguente formula:

$$\frac{x - \min(x)}{\max(x) - \min(x)}$$

A questo punto si calcolano le medie di questi tre valori di rischio nei tre cluster ottenuti in precedenza con l'algoritmo *K-Medoids*.

Cluster	Rischio	Ricoveri	Amputazioni
Cluster 1	0.715	0.747	0.453
Cluster 2	0.345	0.363	0.222
Cluster 3	0.061	0.070	0.027

Tabella 3: *Valori medi nei tre cluster delle variabili RSK, RCV' e AMP' normalizzate*

La Tabella 3 mostra come nel *Cluster 1*, contenente 8 sottosettori, siano raggruppate le osservazioni con incidenze di ricoveri e di mutilazioni più alte, ad altresì il rischio più alto in media. La maggior parte dei sottosettori inclusi in tale cluster appartengono all'industria manifatturiera e solamente il 3% dei lavoratori americani ha un'occupazione in uno di essi. Tutti e cinque i sottosettori a cui è associato il rischio maggiore appartengono al primo cluster. In particolare, il sottosettore più pericoloso in assoluto appartenente a questo cluster è 321 - *Wood product manufacturing*, che presenta il valore più alto sia per *Rischio* sia per *Amputazione*. Per quanto riguarda *Ricovero*, il sottosettore più a rischio è 213 - *Support activities for mining*. Per questi motivi questo cluster può essere visto come quello che contiene i sottosettori **ad alto rischio**.

Settore	N. infortuni	N. dipendenti
Wood product manufacturing	565	378474
Primary metal manufacturing	549	398566
Support activities for mining	563	416230
Support activities for agriculture and forestry	105	96466
Nonmetallic mineral product manufacturing	357	369859

Tabella 4: *I 5 settori con il più alto rischio associato*

Il *Cluster 3* è caratterizzato da valori estremamente bassi in tutte le misure considerate. È il cluster più numeroso con 47 sottosettori, i quali fanno parte nella loro quasi totalità all'ambito dei servizi. Il sottosettore che presenta valori di *Rischio* e di *Ricovero* più bassi in assoluto è 551 - *Management of companies and enterprises*. Con il valore più basso di *Amputazione*, si hanno con un livello pressoché nullo 483 - *Water transportation* e 524 - *Insurance carriers and related*

activities. Questo cluster è quindi costituito dai sottosettori **a basso rischio**.

Settore	N. infortuni	N. dipendenti
Management of companies	12	3308759
Securities, commodity contracts	7	896135
Credit intermediation	31	2760714
Insurance carriers	30	2453404
Religious, grantmaking civic, professional	44	2758413

Tabella 5: *I 5 settori con il più basso rischio associato*

Il *Cluster 2*, composto da 24 sottosettori, presenta livelli intermedi per ognuno dei tre valori in analisi. Pertanto questo gruppo può esser denominato come quello dei sottosettori **a medio rischio**.

6.2 Regressione lineare tra *Paghe* e *Casi*

A questo punto ci si chiede se sussista una corrispondenza - positiva o negativa - tra le variabili *Rischio* e *Paghe* (P): in altre parole, se un'alta pericolosità è remunerata con un alto stipendio, oppure no.

Il modello che si applica è una regressione lineare semplice. La variabile *Paghe* viene divisa per il numero di impiegati e moltiplicata per mille (infatti i valori sono in migliaia di dollari):

$$P' = \frac{1000\ P}{I}$$

dopodiché viene trasformata con un logaritmo allo scopo di ridurre la forte asimmetria positiva che presenta. La specificazione del modello è dunque la seguente:

$$\log P'_i = \beta_0 + \beta_1\ RSK_i + \mathcal{E}_i \qquad i = 1, \ldots, 79$$

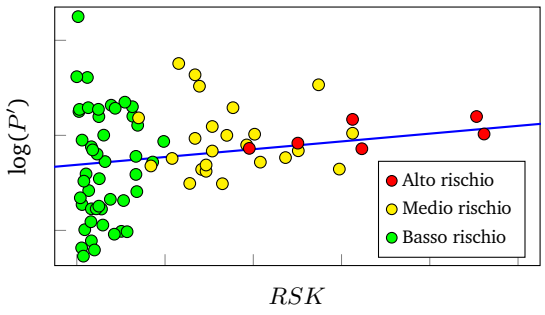

Figura 5: *Grafico a dispersione tra la variabile* Rischio *e il logaritmo di Paghe; i punti sono colorati in base al cluster di appartenenza. In blu è riportata la retta di regressione.*

Con i dati si ottiene il modello stimato:

$$\widehat{\log P'}_i = \widehat{\beta}_0 + \widehat{\beta}_1\ RSK_i = 10.69 + 0.41\ RSK_i$$

Il coefficiente $\widehat{\beta}_1$ ottenuto è associato ad un pvalue pari a 0.096 e quindi risulta essere appena significativo soltanto per un livello di significatività piuttosto alto ($\alpha = 0.1$). Nella Figura 5 si può osservare l'assenza di qualsivoglia correlazione lineare, e graficamente non è possibile identificare se la relazione tra le due variabili sia positiva o negativa. Una spiegazione plausibile a questo fatto è che le professioni del settore terziario non di rado richiedono alte qualifiche (quindi sono ben ricompensate) e hanno una probabilità di infortuni sul lavoro praticamente nulla.

6.3 Analisi delle componenti principali

Lo studio delle correlazioni tra le variabili standardizzate contenute nella matrice $\mathbf{X}$ suggerisce un approccio diverso, che riesca a cogliere e a mettere in risalto quei fattori che altrimenti resterebbero nascosti. Infatti la matrice $\mathbf{R}$ delle correlazioni denota un numero elevato di variabili con un legame molto forte, le quali nel processo di *clustering* mettono in ombra altre variabili potenzialmente rilevanti. La tecnica di cui si fa uso è l'Analisi delle componenti principali (PCA) effettuata sulla matrice $\mathbf{X}$, dalla quale vengono estromesse con un nodo *Correlation filter* le variabili con correlazione molto alta (maggiore di 0.80), in quanto apportano informazioni già presenti in altre variabili. Denotando con p il numero delle colonne di X (ridotto a 10), le componenti principali $\mathbf{Y}_1, \ldots, \mathbf{Y}_p$ (PC) sono combinazioni lineari della matrice $\mathbf{X}$, non correlate tra loro e ordinate in senso decrescente rispetto alle loro varianze $\lambda_1, \ldots, \lambda_p$ (le quali corrispondono agli autovalori della matrice $\mathbf{R}$): la prima PC sarà quella variabile artificiale in grado di ricostruire la massima varianza del set di variabili $\mathbf{X}_1, \ldots, \mathbf{X}_p$ prese in considerazione. I valori che si generano dalla combinazione lineare sono detti punteggi di ogni componente principale. Seppure le PC siano p, ovvero pari al numero delle variabili $\mathbf{X}_j$ di partenza, per comprendere meglio il fenomeno in analisi e per ragioni di sintesi è utile sostituire le p variabili $\mathbf{X}_j$ tra loro correlate con un numero ridotto di PC che siano di facile interpretazione.[7]

Non c'è un metodo univoco per stabilire il numero preciso di componenti principali da conservare. Di seguito se ne elencano due:

- scree-plot: grafico degli autovalori in funzione del numero di PC; poiché gli autovalori sono decrescenti, il grafico assume l'aspetto di una spezzata con pendenza negativa. Una variazione di pendenza significativa indica il numero di componenti da tenere in considerazione.
- quota di varianza totale spiegata: si scelgono le componenti principali che tengono conto di una quota sufficientemente elevata di varianza totale spiegata. Estraendo le PC a partire dalla matrice $\mathbf{R}$ si terranno in considerazione pertanto solo quelle PC con varianza (ovvero autovalore associato) maggiore di 1.

Nel caso in esame si applica il secondo criterio, in quanto come si può notare dalla Figura 6, tra il primo e il secondo autovalore si registra una variazione di pendenza notevole, tuttavia considerare una componente significherebbe tenere soltanto l'informazione fornita dal blocco di variabili più correlate. Pertanto, il numero di PC tenuto in considerazione è 3, cioè quelle a cui è associato un autovalore maggio-

[7] Per maggiori dettagli sull'Analisi delle componenti principali si rimanda al paper [8]

re di 1; la percentuale di varianza assorbita da queste tre componenti è circa il 72%.

Come detto, quindi, la prima PC riassume gran parte della variabilità (50% circa) e delle informazioni delle variabili; di conseguenza essa è molto correlata con le variabili. Fattori degni di nota emergono invece studiando le correlazioni delle altre due PC con le variabili input dell'algoritmo.

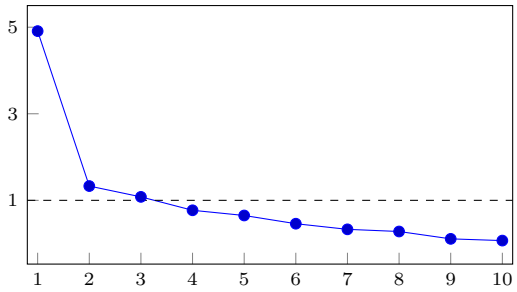

Figura 6: *Scree-plot degli autovalori connessi alle componenti principali.*

Come si può notare nella Tabella 6, la prima componente presenta forti legami con le variabili N_1 (infortuni traumatici), N_2 (ferite aperte) e S_4 (incidenti avvenuti maneggiando materiali, macchinari o altri strumenti industriali). La seconda PC invece è molto correlata positivamente con E_1 (violenza) e S_5 (incidenti in cui sono implicati esseri viventi), negativamente con N_3 (ustioni) e E_3 (esplosioni e incendi). La terza PC a sua volta ha le correlazioni più forti con E_3, S_5 e E_2 (incidenti avvenuti durante fasi di trasporto). Una prima possibile interpretazione è la seguente: la prima componente ingloba al suo interno tutti gli infortuni "generici", come ad esempio cadute, amputazioni e altri incidenti avvenuti in modo fortuito; la seconda mette in rilievo infortuni dovuti a episodi violenti, quali ad esempio furti e rapine, inoltre tiene in considerazione situazioni molto gravi come incendi o esplosioni; la terza, infine, oltre ad aggregare i fattori già detti, tiene conto del rilevante numero di infortuni avvenuti durante il trasporto di oggetti, quali ad esempio materie prime o merce in consegna.

PC	N_1	N_2	N_3	N_5	E_1	E_2	E_3	E_7	S_4	S_5
PC1	**0.94**	0.76	0.7	0.74	0.42	0.64	0.7	0.65	**0.82**	0.47
PC2	-0.06	-0.23	-0.39	0.27	**0.74**	0.19	-0.35	0.27	-0.31	**0.42**
PC3	-0.13	-0.14	0.28	-0.03	0.28	-0.44	**0.46**	-0.44	0.07	**0.52**

Tabella 6: *Correlazioni lineari tra le PC e le variabili contenute nella matrice* $\mathbf{X}$. *In grassetto evidenziati i due valori più alti in ogni riga.*

6.3.1 Clustering sui punteggi delle componenti

A questo punto è possibile effettuare *clustering* sui punteggi delle prime tre componenti, il cui apporto è dunque più proporzionato, volto a dare maggior peso a eventi la cui gravità è indiscutibilmente superiore. L'algoritmo scelto è di tipo gerarchico, in particolare si utilizza il metodo di Ward.

Dal dendrogramma (Figura 7) emergono quattro clusters ben definiti.

- Procedendo da destra verso sinistra, il primo racchiude dei sottosettori con caratteristiche simili: ad esempio 211 - *Oil and gas extraction*, 213 - *Support activities for mining*, 324 - *Petroleum and coal products manufacturing* e 331 - *Primary metal manufacturing* fanno parte di questo gruppo; questi sottosettori, oltre a presentare un livello di *Rischio* molto preoccupante, hanno valori della variabile E_3 (incendi ed esplosioni) estremamente alti.
- Il secondo da destra comprende sottosettori nei quali non si denota un tasso di *Rischio* molto alto, ma sono avvenuti molti episodi di violenza o crimine. Sono in questo cluster ad esempio *Couriers and messengers*, 115 - *Support activities for agriculture and forestry* e 712 - *Museums, historical sites, and similar institutions*.
- Nel terzo cluster da destra si trovano quei sottosettori con un livello della variabile *Rischio* molto alto, per via del fatto che si sono verificati innumerevoli incidenti, soprattutto dovuti a comportamenti incauti dei lavoratori, oppure a fattori semplicemente fortuiti. Fanno parte di questo gruppo 321 - *Wood product manufacturing* e molti altri sottosettori appartenenti all'industria manifatturiera.
- Si trovano infine nel quarto ed ultimo cluster i sottosettori meno pericolosi, i quali presentano pochissime situazioni spiacevoli.

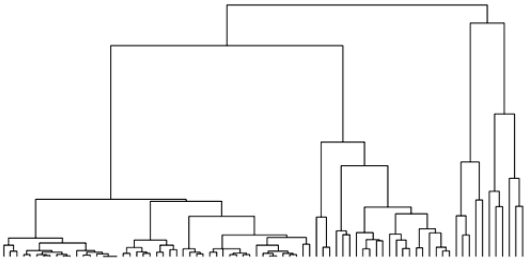

Figura 7: *Dendrogramma relativo ai cluster ottenuti a partire dalle PCA*

I risultati del raggruppamento con il metodo di Ward forniscono molti spunti interessanti, raggiunti solo grazie all'output di un'Analisi delle componenti principali. Effettuare *clustering* su tutte le variabili di partenza produce una suddivisione molto simile a considerare soltanto la prima componente principale; pertanto, è evidente che l'introduzione delle due componenti porta alla scoperta di alcuni fattori che altrimenti resterebbero occulti.

7 Criticità e limiti

Le variabili di cui si dispone permettono di conoscere i dettagli dell'incidente e del tipo di trauma subito dal malcapitato; tuttavia, nel dataset emphSeverelyInjuredWorkers sono inclusi solo episodi gravi che hanno comportato il ricovero o un'amputazione ma non la morte. Per un'analisi ancora più dettagliata e completa sarebbe utile disporre di tutti gli incidenti avvenuti in ciascun settore, da quelli meno gravi a quelli che hanno causato il decesso del lavoratore. In

questo modo si potrebbero individuare i settori che hanno un maggiore tasso di mortalità ed eventualmente suggerire l'impiego di ulteriori misure di sicurezza a tutela dei dipendenti. Inoltre il dataset risulta sprovvisto di informazioni relative alle caratteristiche dei lavoratori come ad esempio l'età o il genere.

Conoscere il periodo di permanenza in ospedale del paziente potrebbe essere rilevante per attribuire un diverso livello di gravità ad ogni infortunio. Con i dati attualmente disponibili non è possibile distinguere infortuni che portano a lunghe degenze da episodi che comportano un semplice day-hospital. Sapere se dopo l'incidente il soggetto abbia ripreso a lavorare e ad eseguire le proprie mansioni permetterebbe di valutare l'impatto economico e sociale degli infortuni.

8 Conclusioni

La cluster analysis ha suddiviso gli incidenti in tre gruppi con composizioni molto diverse: differiscono tra di loro per tipologia di settore predominante, rischio di incidente, rischio di mutilazione e di infortunio. Queste informazioni potrebbero essere impiegate dal governo americano o da un ente come l'OSHA al fine di adottare ulteriori misure di sicurezza e quindi garantire un ambiente di lavoro sicuro, soprattutto in quei settori che appartengono al cluster più pericoloso.

Inoltre è emerso che solamente una bassa percentuale di americani, il 3%, lavori in settori appartenenti a questo cluster e quindi con fattore di rischio molto alto. Questo rischio non ha una adeguata remunerazione, in quanto non esiste alcuna correlazione tra rischio e stipendio medio del sottosettore. Ciò conferma come in una società industriale contemporanea, quale quella nordamericana, il settore trainante sia il terziario, che è caratterizzato da lavori nell'ambito dei servizi con bassa incidenza di infortuni.

Infine l'analisi delle componenti principali consente di raggruppare i settori in modo ancora più dettagliato. Si potrebbe pensare di adottare misure di precauzione appositamente per i settori soggetti a rischio di esplosione, altre ancora per i settori caratterizzati da episodi di violenza o crimini.

Appendice A

Per maggior chiarezza vengono riportate in questa appendice le variabili utilizzate nel primo dataset. Per l'identificazione dei codici di Nature, PartofBody, Event e Source si rimanda alla fonte disponibile sulla piattaforma Kaggle[4].

SEVERE INJURY

- **EventDate**. Data dell'infortunio
- **Employer**. Nome dell'azienda in cui è avvenuto il caso
- **City**. Città in cui ha sede l'azienda
- **State**. Stato USA in cui ha sede l'azienda (sigla)
- **MacroState**. Area geografica USA in cui opera l'azienda
- **Zip**. ZIP code a cinque cifre, codice postale americano
- **Latitude**. Latitudine geografica
- **Longitude**. Longitudine geografica
- **PrimaryNAICS**. Codifica NAICS dell'azienda a sei livelli di dettaglio (sei cifre), identifica l'industria in cui opera lo stabilimento
- **Hospitalized**. Ricovero in ospedale (dummy). 1 = ricoverato, 0 = non ricoverato
- **Amputation**. Amputazione di una parte del corpo (dummy). 1 = amputato, 0 = non amputato.
- **Nature**. Classificazione della natura dell'evento, ad esempio eventi traumatici, ferite aperte, malattie. Codifica OIICS a quattro livelli di dettaglio.
- **PartofBody**. Classificazione della parte del corpo lesa, ad esempio nuca, collo, arti superiori o inferiori. Codifica OIICS a quattro livelli di dettaglio.
- **Event**. Classificazione dell'evento da cui è scaturito l'infortunio, ad esempio scivolate o cadute, reazioni a sostanze, casi di violenza, incidenti su mezzi di trasporto. Codifica OIICS a quattro livelli di dettaglio.
- **Source**. Classificazione dell'oggetto coinvolto nell'incidente, ad esempio veicoli, animali, macchinari, strumenti. Codifica OIICS a quattro livelli di dettaglio.

Appendice B

Per maggiore chiarezza vengono riportate in questa appendice le variabili utilizzate nel secondo dataset.

NAICS

- **Naics**. Codice settoriale Naics a tre livelli di dettaglio (tre cifre)
- **Nome**. Nome identificativo del settore
- **Stabilimenti**. Numero di imprese per sottosettore
- **Paghe**. Retribuzioni annuali degli impiegati (in migliaia di dollari) totali per sottosettore
- **Impiegati**. Numero di dipendenti per settore

Riferimenti

[1] *Occupational Safety and Health Administration (OSHA)*. `https://www.osha.gov/about.html`.

[2] *Occupational Injury and Illness Classification System (OIICS)*. `https://wwwn.cdc.gov/wisards/oiics/`.

[3] *NAICS*. `https://www.census.gov/eos/www/naics/`.

[4] *Kaggle, Severely Injured Workers*. `https://www.kaggle.com/jboysen/injured-workers`.

[5] *KNIME*. `https://www.knime.com/`.

[6] *RStudio*. `https://www.rstudio.com/`.

[7] *clValid package*. `http://cran.us.r-project.org/web/packages/clValid/vignettes/clValid.pdf`.

[8] Svante Wold, Kim Esbensen e Paul Geladi. «Principal component analysis». In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.