

Non-life insurance: analisi del danno per sinistro

Carbonera Francesco
mat. EC7100326, Università degli Studi di Trieste
francesco.carbonera@studenti.units.it

Abstract

L'analisi e l'utilizzo di tecniche modellistiche all'interno del ramo assicurativo non-life hanno un ruolo determinante. Mentre dal punto di vista esplorativo è di fondamentale importanza analizzare l'utilità delle variabili, dalla parte modellistica è fondamentale la necessità di dover determinare un modello capace non solo di adattarsi bene ai dati disponibili ma di avere anche una buona capacità previsionale. Nel seguente elaborato si analizzerà un dataset contenente un discreto numero di polizze con lo scopo di stimare dei GLM per la determinazione della variabile *danno* per sinistro.

Indice

1. Introduzione
 - 1.1 Mission statement
2. Il dataset Polizze
3. Analisi esplorativa
 - 3.1. Relazione tra variabili esplicative
4. Modellizzazione con l'uso dei GLM
 - 4.1. Modello Gamma con dati individuali
 - 4.2. Modello Gamma con dati raggruppati
 - 4.2.1. Modello Gamma con dati raggruppati e dispersione con stima di Pearson
 - 4.2.2. Modello Gamma con dati raggruppati e dispersione con stima di quasi-devianza
5. Conclusioni

1.Introduzione

Nel 2019, le imprese assicuratrici hanno rimborsato costi per sinistri per 96 mld¹, di questi, 76 mld di euro sono derivati dai rami vita mentre per la parte rimanente, 20 mld, dai rami danni. Inoltre, per i rami vita i costi rimborsati per sinistri ammontano al 71% dei premi raccolti, meno nei rami danni dove si attestano al 64%.

Nel 2019 per il ramo RC Auto e natanti, il costo medio dei sinistri risarciti nello stesso anno di generazione è stato di 2.611 euro². Il risultato è rimasto invariato rispetto all'anno precedente, con una variazione positiva di 32 euro. Anche nel periodo 2014-2019, l'incremento è stato minimo (solo del 3,7% a prezzi costanti).

A livello complessivo, tenendo conto anche dei sinistri posti a riserva ma non ancora liquidati, per la generazione del 2019, il costo medio ammontava a 4.348 euro. L'importo è rimasto stabile rispetto all'anno prima, mentre nell'arco dei 5 anni è diminuito del 6,4%.

Anni	Costo medio dei sinistri pagati	Costo medio dei sinistri riservati		Costo medio complessivo dei sinistri		Premio puro ^(a)	
		Al netto della stima IBNR	Al lordo della stima IBNR	Al netto della stima IBNR	Al lordo della stima IBNR	Valore	Var. (%) ^(b)
<i>Valori a prezzi costanti 2019^(c)</i>							
2014	2.517	9.758	8.677	4.647	4.758	288	-0,5%
2015	2.526	9.817	8.702	4.587	4.701	286	-0,8%
2016	2.544	9.604	8.489	4.495	4.593	284	-2,9%
2017	2.557	9.332	8.360	4.396	4.504	276	-3,4%
2018	2.579	9.567	8.469	4.384	4.480	266	-1,1%
2019	2.611	9.582	8.551	4.348	4.469	263	
Variazione 2019/2014	+3,7%	-1,8%	-1,5%	-6,4%	-6,1%		-8,6%
<i>Valori a prezzi correnti</i>							
2014	2.455	9.758	8.677	4.532	4.641	281	-3,2%
2015	2.460	9.817	8.702	4.467	4.578	279	-0,7%
2016	2.476	9.604	8.489	4.374	4.469	276	-0,8%
2017	2.516	9.332	8.360	4.326	4.432	271	-1,9%
2018	2.566	9.567	8.469	4.361	4.457	265	-2,3%
2019	2.610	9.582	8.851	4.348	4.469	263	-0,7%
Variazione 2019/2014	+6,3%	-1,8%	-1,5%	-4,1%	-3,7%		-6,3%

(a) Prodotto tra frequenza dei sinistri denunciati (tav. 3) e costo medio complessivo dei sinistri, entrambi al lordo della stima IBNR. – (b) Variazioni rispetto all'anno precedente. – (c) Deflatore utilizzato: indice dei prezzi al consumo per famiglie di operai e di impiegati (FOI) al netto dei tabacchi.

Figure 1: *Indicatori di costo medio dei sinistri denunciati nell'anno di accadimento fonte:IVASS,2020*

¹ I principali numeri delle assicurazioni in Italia - 2019, IVASS

² Bollettino Statistico Anno VII - n. 15 - dicembre 2020, IVASS

A fronte di questi importi monetari, diventa essenziale poter disporre, con l'utilizzo di una base dati consistente e l'ausilio di tecniche statistiche e sotto opportune ipotesi matematiche³, di alcuni modelli che siano in grado di raccogliere e sintetizzare informazioni e relazioni. La finalità è duplice: sia di disporre di una valutazione utile per il pricing nella fase a priori, sia di costruire una base tecnica di partenza (e tutto ciò che concerne valutazioni e stime di riserve).

1.1. Mission statement

In questo elaborato verranno studiati dei modelli utili a descrivere la variabile che rappresenta il danno medio per sinistro per polizza. Nella prima parte si effettua una rapida rappresentazione delle variabili esplicative rispetto alla variabile dipendente. In particolare, si utilizzeranno le stesse variabili, con la stessa fattorizzazione, utilizzate nel *modello di Poisson con dati raggruppati* per la stima del *numero di sinistri* per polizza. Nella seconda parte verranno analizzati dei modelli GLM, in particolare il modello con distribuzione di Gamma con dati individuali e con dati raggruppati. L'elaborato viene svolto principalmente con l'utilizzo del software SAS. Tuttavia per alcune tecniche e grafici si preferisce usare il software RStudio.

³ Al fine di poter accogliere l'ipotesi di distribuzione composta del numero aleatorio X (numero aleatorio che rappresenta il risarcimento totale), consistente in termine di valori attesi a $E(X) = E(N)E(Y)$, è necessario accogliere:

- per ogni $n > 0$, $Y_1|N = n, \dots, Y_n|N = n$ sono stocasticamente indipendenti ed identicamente distribuiti,
- la distribuzione di probabilità di $Y_i|N = n$, $i \leq n$ non dipende da n .

2. Il dataset Polizze

L'elaborato consiste nell'analisi del dataset Polizze. Il dataset contentente 11748 polizze RCAuto sinistrate e per ognuna si dispone di 11 variabili. Per il problema qui trattato la variabile *nsin* non viene considerata, quindi per il momento è possibile eliminarla insieme anche all'*esposizione*. Inoltre, a livello funzionale si trasforma il valore "NA" della variabile *Provincia*, che rappresenta le polizze nelle quali gli assicurati hanno residenza in provincia di Napoli, in "NAP". Questo serve per evitare che il valore venga scambiato per un "not available". Le variabili contenute nel dataset sono le seguenti:

- *Sesso*, variabile fattoriale (M/F) che rappresenta il sesso dell'assicurato;
- *Eta*, variabile numerica che rappresenta l'età dell'assicurato;
- *Provincia*, variabile fattoriale che rappresenta la provincia dell'assicurato;
- *Capoluogo*, variabile fattoriale (SI/NO) che serve ad identificare se l'assicurato abita nel capoluogo di provincia;
- *Bendie*; variabile fattoriale (B/D) che rappresenta la tipologia di alimentazione del veicolo dell'assicurato, i cui valori possibili sono *benzina* o *diesel*;
- *Potkil*, variabile numerica che rappresenta la potenza in kilowatt del veicolo dell'assicurato;
- *Massa*, variabile numerica che rappresenta la massa del veicolo dell'assicurato;
- *Potf*, variabile numerica che rappresenta la potenza fiscale del veicolo dell'assicurato;
- *Dannototale*, variabile numerica che rappresenta il danno totale per polizza;
- *nsin*, variabile numerica che rappresenta il numero di sinistri per polizza;
- *Dannomedio*, variabile numerica che rappresenta il rapporto tra il danno totale e il numero di sinistri per singola polizza.

Alcune variabili sono state ricodificate per essere raggruppate in clusters. In particolare, si osservano queste variabili:

- *leveleta*, variabile fattoriale per la variabile *Eta*, i cui livelli sono 18-20, 21-26, 27-43, 44-58, 59-69, 70-81 e 82-;
- *levelpotf*, variabile fattoriale per la variabile *Potf*, i cui livelli sono -14, 15-18, 19-21 e 22-;
- *cluster*, variabile fattoriale per la variabile *Provincia*, i cui livelli raggruppano le diverse province italiane.

Il dataset è quindi così formato:

```
'data.frame': 11748 obs. of 9 variables:
 $ Sesso      : Factor w/ 2 levels "F","M": 2 1 2 2 2 2 2 2 2 1 ...
 $ Capoluogo : Factor w/ 2 levels "NO","SI": 1 1 1 1 1 1 2 1 1 1 ...
 $ leveleta   : Factor w/ 7 levels "18-20","21-26",...: 2 4 2 2 6 3 3 4 4 3 ...
 $ levelpotf  : Factor w/ 4 levels "-14","15-18",...: 3 3 2 1 1 2 2 2 2 1 ...
 $ cluster    : Factor w/ 15 levels "1","2","3","4",...: 7 7 7 7 7 7 7 7 7 7 ...
 $ Bendie     : Factor w/ 2 levels "B","D": 2 2 2 1 1 2 1 1 2 1 ...
 $ nsin       : int 1 1 1 1 1 2 1 1 1 ...
 $ Dannotentale: num 3070 763 1032 987 1618 ...
 $ Dannomedio : num 3070 763 1032 987 1618 ...
```

Le prime 5 osservazioni sono le seguenti:

Oss	Sesso	Eta	Prov	Capoluogo	Bendie	Potkil	Massa	potf	nsin	dannotentale	dannomedio	leveleta	levelpotf	Cluster
1	M	55	AR	NO	B	80	1180	18	1	359.06	359.06	44-58	15-18	9
2	M	56	TE	NO	B	55	1120	17	1	1617.55	1617.55	44-58	15-18	9
3	F	32	PD	NO	B	25	700	10	2	1914.67	957.34	27-43	-14	12
4	F	33	VC	NO	B	36	795	14	1	179.53	179.53	27-43	-14	11
5	F	56	BL	NO	B	66	1050	17	1	691.19	691.19	44-58	15-18	12

3. Analisi esplorativa

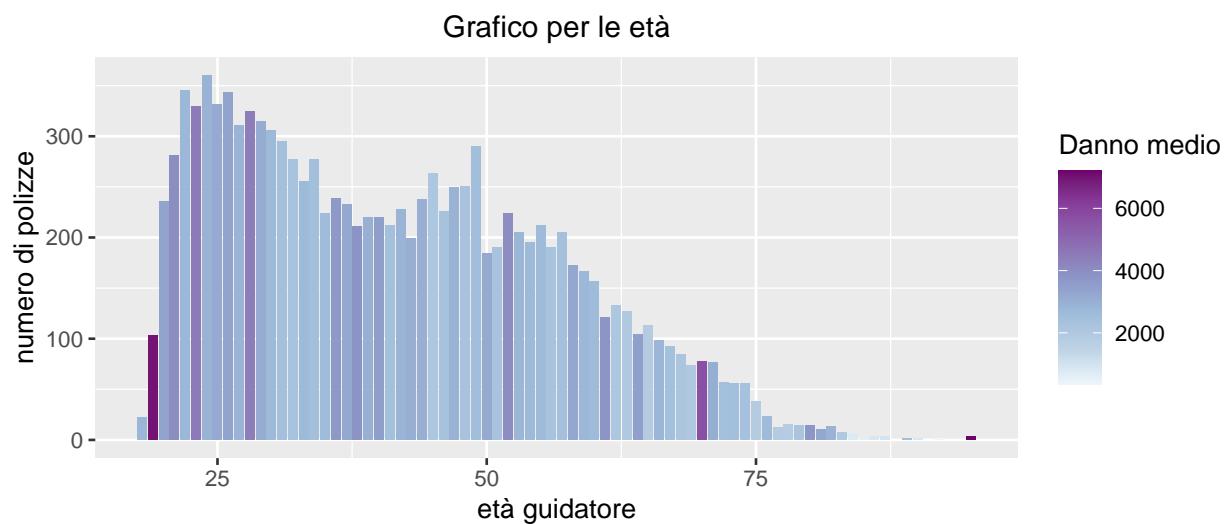
Si osserva velocemente come le variabili esplicative siano distribuite rispetto al **Dannomedio** alla luce delle fattorizzazioni effettuate. In generale, questa è la composizione del database per ogni variabile.

Sesso	Capoluogo	leveleta	levelpotf	cluster	Bendie
F:3604	NO:9126	18-20: 361	-14 :5454	2	:2395 B:10492
M:8144	SI:2622	21-26:1989	15-18:4588	3	:1883 D: 1256
		27-43:4347	19-21:1388	6	:1736
		44-58:3293	22- : 318	12	:1280
		59-69:1269		9	:1050
		70-81: 450		11	: 859
		82- : 39			(Other):2545
nsin		Dannotentale	Dannomedio		
Min. :1.00	Min. : 0.9	Min. : 0.9	Min. : 0.9		
1st Qu.:1.00	1st Qu.: 763.0	1st Qu.: 740.6	1st Qu.: 740.6		
Median :1.00	Median : 1617.6	Median : 1617.6	Median : 1617.6		
Mean :1.08	Mean : 3240.8	Mean : 3035.3	Mean : 3035.3		
3rd Qu.:1.00	3rd Qu.: 2545.7	3rd Qu.: 2545.7	3rd Qu.: 2545.7		
Max. :5.00	Max. :452412.0	Max. :452412.0	Max. :452412.0		

Questi sono alcuni grafici per analizzare meglio come sia formato il dataset:

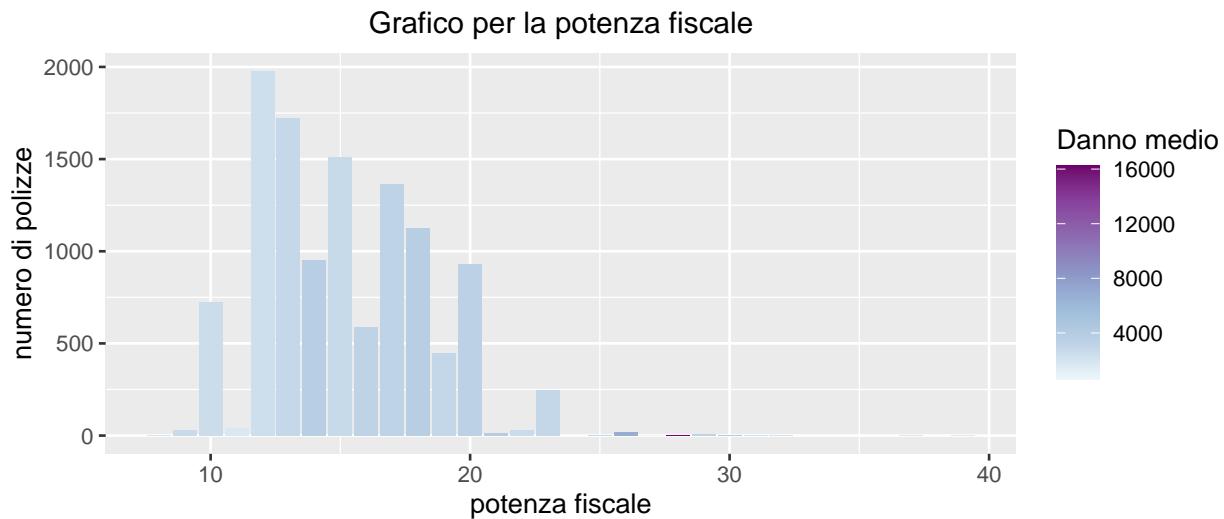
- *età*

Si osserva l'istogramma per le varie età evidenziando il numero di polizze e il danno medio. Per età superiori a 75 anni si notano poche polizze sinistre. Per queste età, il valore del **dannomedio** può essere influenzato dalla poca disponibilità di dati.

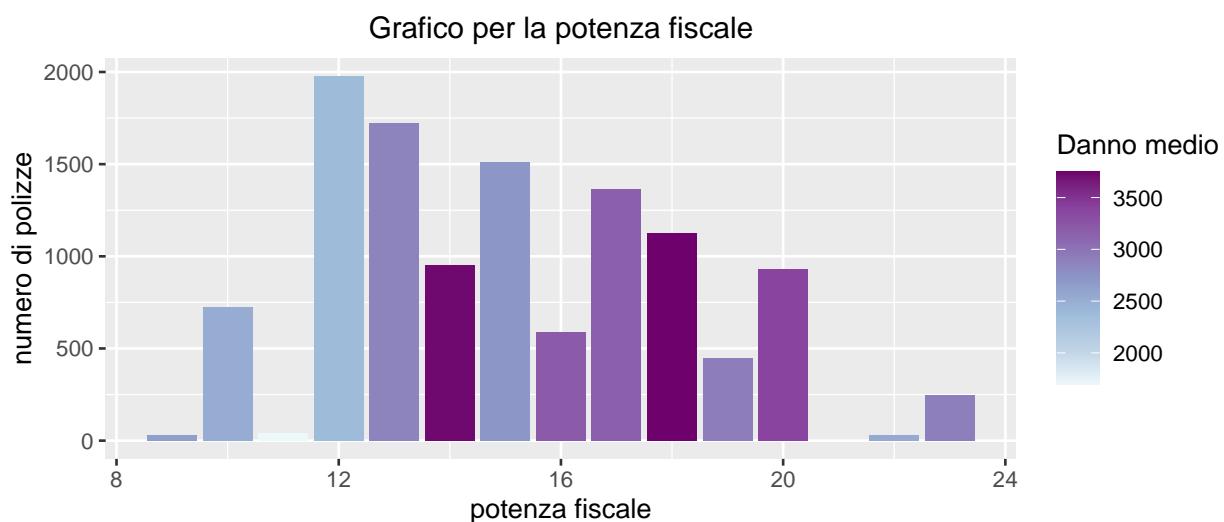


- *potenza fiscale*

Si osserva l'istogramma per i diversi valori della potenza fiscale evidenziando il numero di polizze e il danno medio. Le polizze sinistrate si concentrano tra i valori 10-20. Tuttavia, non c'è diversità di colore tra i vari valori. Insieme al fatto che la scala ha un ampio range, questo problema rivela come ci sia almeno un valore per la potenza fiscale (in questo caso per **Potf** uguale a 28, si osservano due polizze sinistrate con un valore alto di **dannomedio**) con un esiguo numero di osservazioni ma con un alto **Dannototale**.

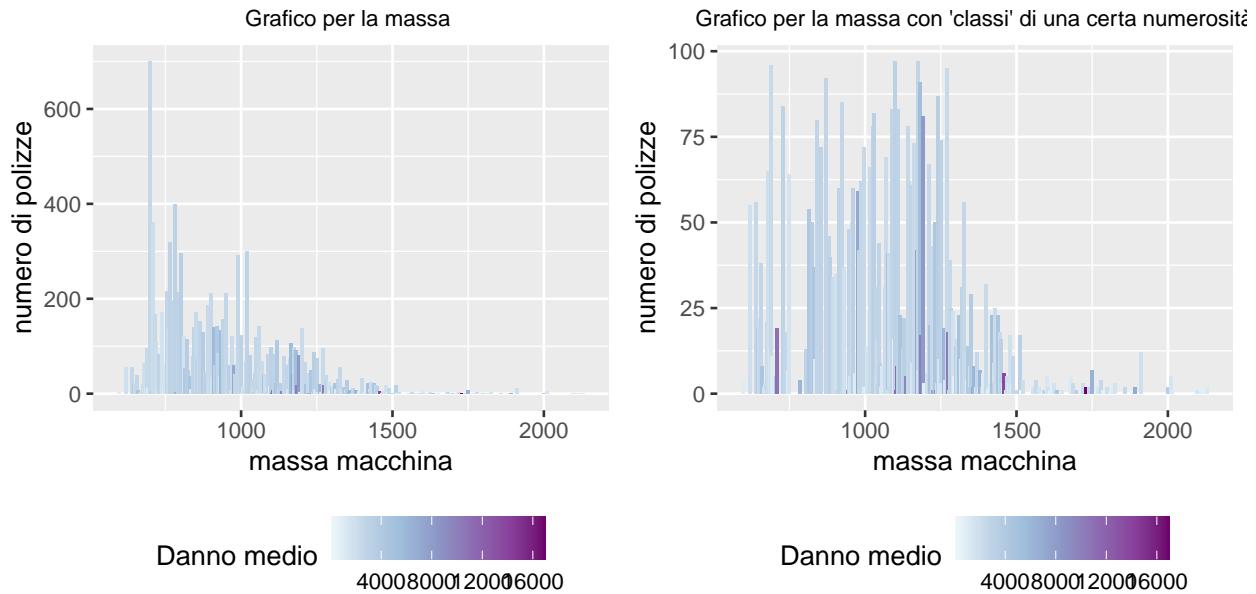


Se si rappresenta in un istogramma i valori della potenza fiscale con numerosità superiore alle 20 polizze, il risultato è questo:



- *massa*

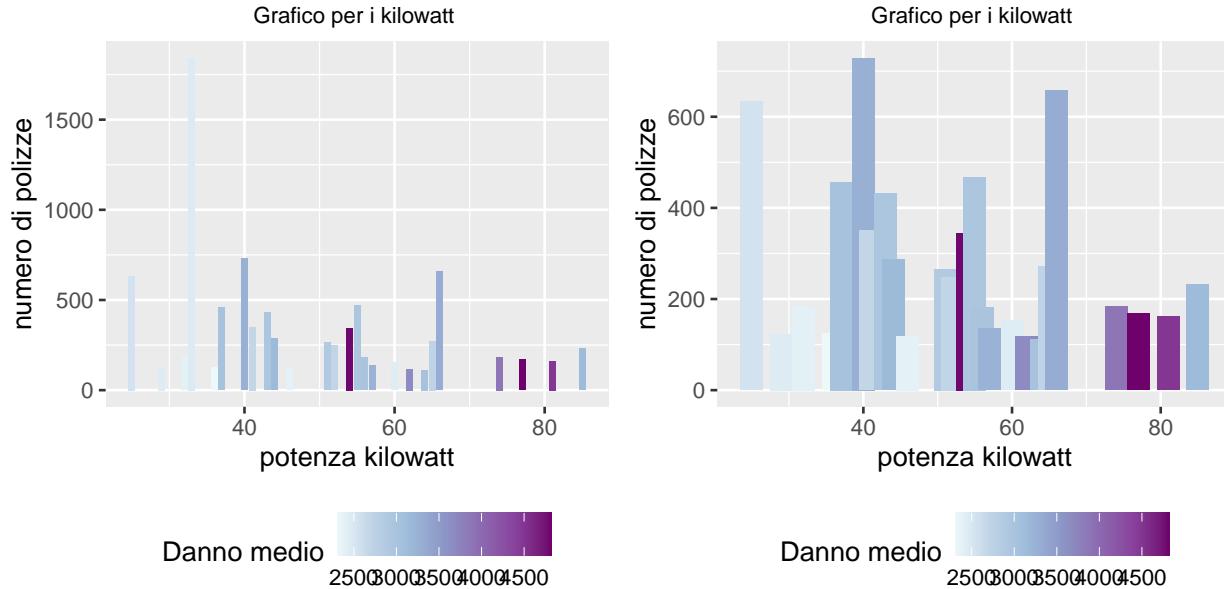
Si osserva se l'istogramma delle polizze per i differenti valori della massa possa risolvere il problema legato al raggruppamento. Infatti, per il problema del *numero di sinistri* non è stato possibile effettuare il raggruppamento a causa della grande variazione della frequenza sinistri ma anche per ragioni tecniche come ad esempio il fatto che il raggruppamento avrebbe creato delle problematiche in fase di assegnazione della classe per valori non ancora osservati.



Anche in questo caso, il grafico rivela come visivamente non sia utile effettuare una cluster analysis.

- *potenza in kilowatt*

Si crea l'istogramma per la variabile **Potkil** e per questo vengono effettuate le stesse considerazioni della variabile **Massa**. Anche in questo caso non risulta opportuno effettuare una cluster analysis.

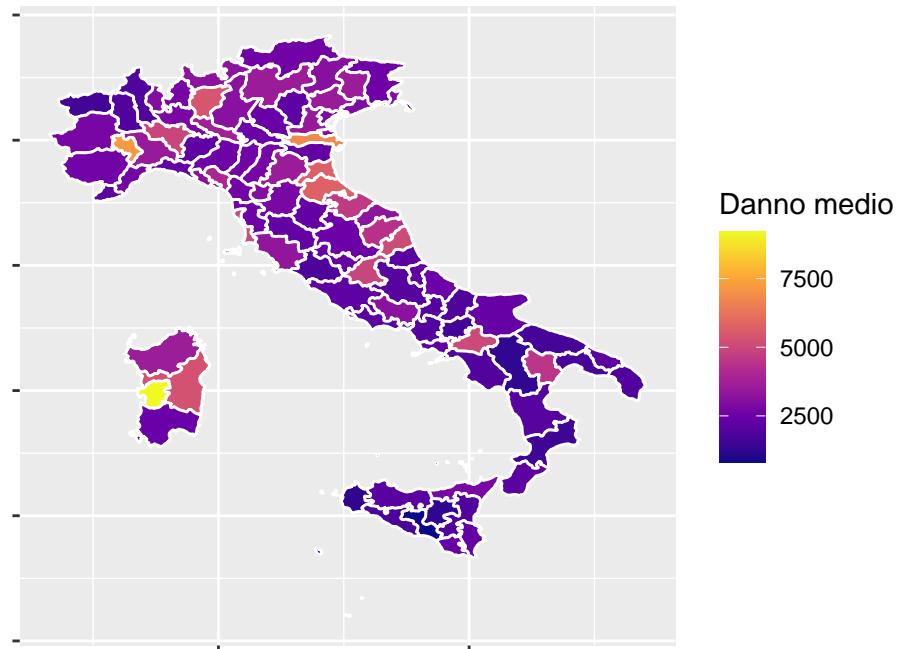


- *province*

Per la variabile **Prov** valgono le stesse considerazioni fatte per la stima del modello per sinisitri. E' una variabile fattoriale con 104 livelli. Si osserva che per le province di *Barletta Andria Trani*, *Fermo*, *Monza e della Brianza* e *Sud Sardegna* non ci sono osservazioni all'interno del dataset e vengono effettuate queste modifiche:

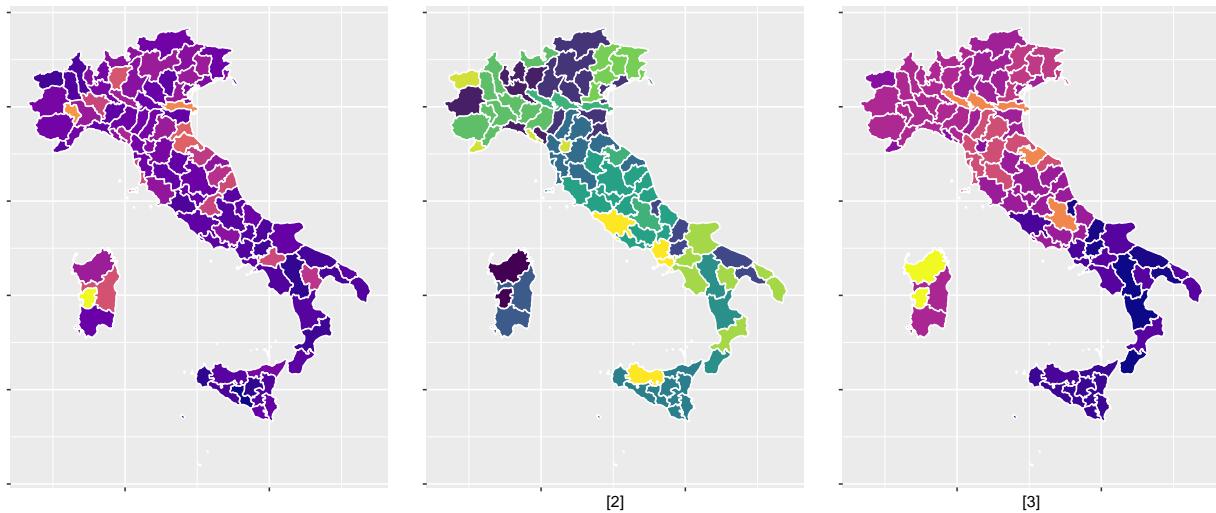
- la provincia di **Biella**, nata dallo scorporo della provincia di **Vercelli**, viene inserita in quest'ultima;
- la provincia di **Lecco** viene inserita nella provincia di **Como**;
- la provincia di **Lodi** viene inserita nella provincia di **Milano**;
- la provincia di **Prato** viene inserita nella provincia di **Firenze**;
- la provincia di **Rimini** viene inserita nella provincia di **Forlì**;
- la provincia di **Verbano** viene inserita nella provincia di **Novara**;
- le province di **Crotone** e **Vibo Valencia** vengono inserite nella provincia di **Catanzaro**;
- viene eliminata l'osservazione che ha per provincia la sigla **RSM** (Repubblica di San Marino);
- dal punto di vista toponomastico la provincia attuale denominata come *Forlì Cesena* prende il nome di *Forlì* (viene ricodificata come **Forli** per evitare problemi legati ai caratteri) e la provincia di *Pesaro Urbino* prende il nome di **Pesaro**.

Il grafico del **dannomedio** per provincia è il seguente:



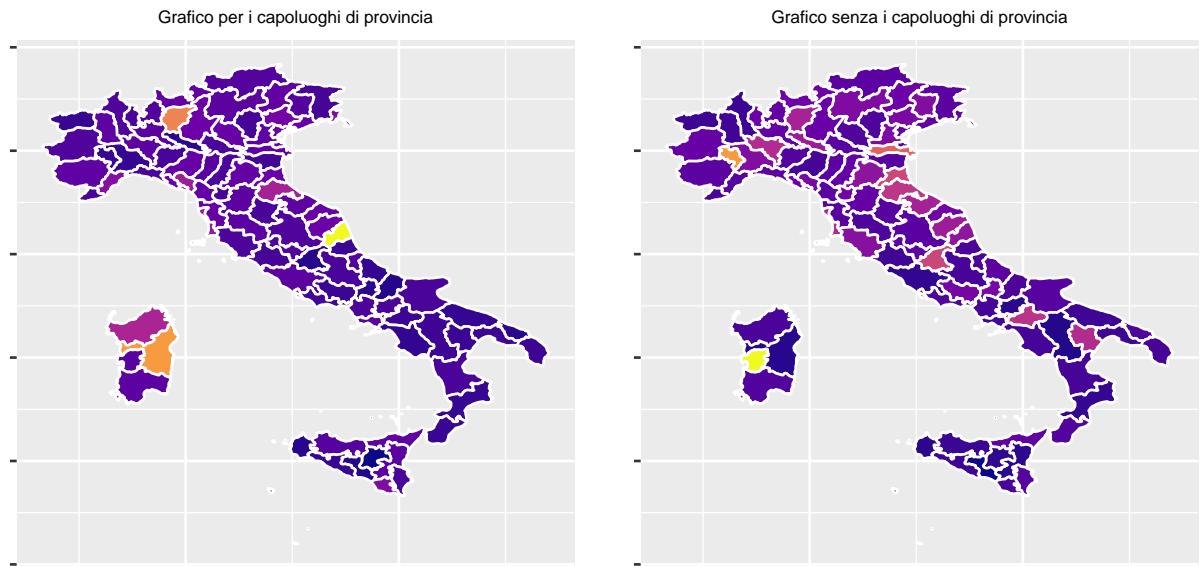
Introducendo la fattorizzazione per livelli della variabile **Provincia** è possibile visualizzare i seguenti grafici.

- [1] le province valutate per **Dannomedio**;
- [2] le province valutate per **cluster** di appartenenza;
- [3] le province valutate per **Dannomedio** calcolata nei **clusters**.

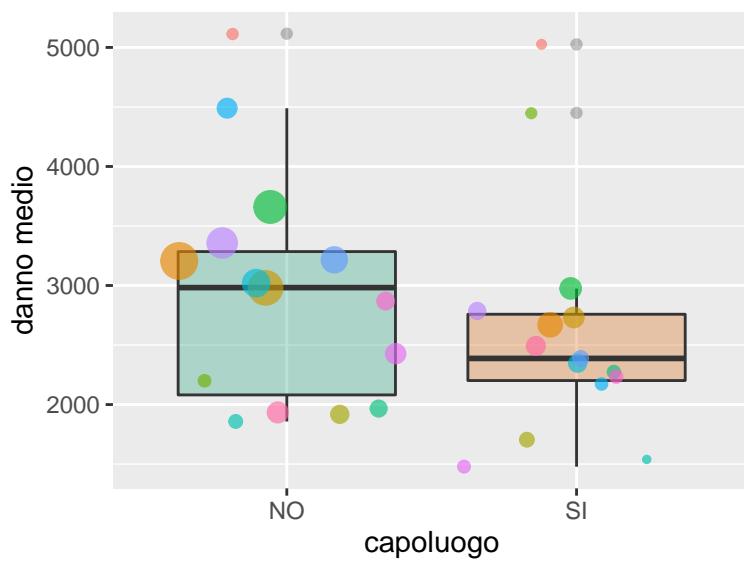


Si mostrano le principali statistiche per le altre variabili.

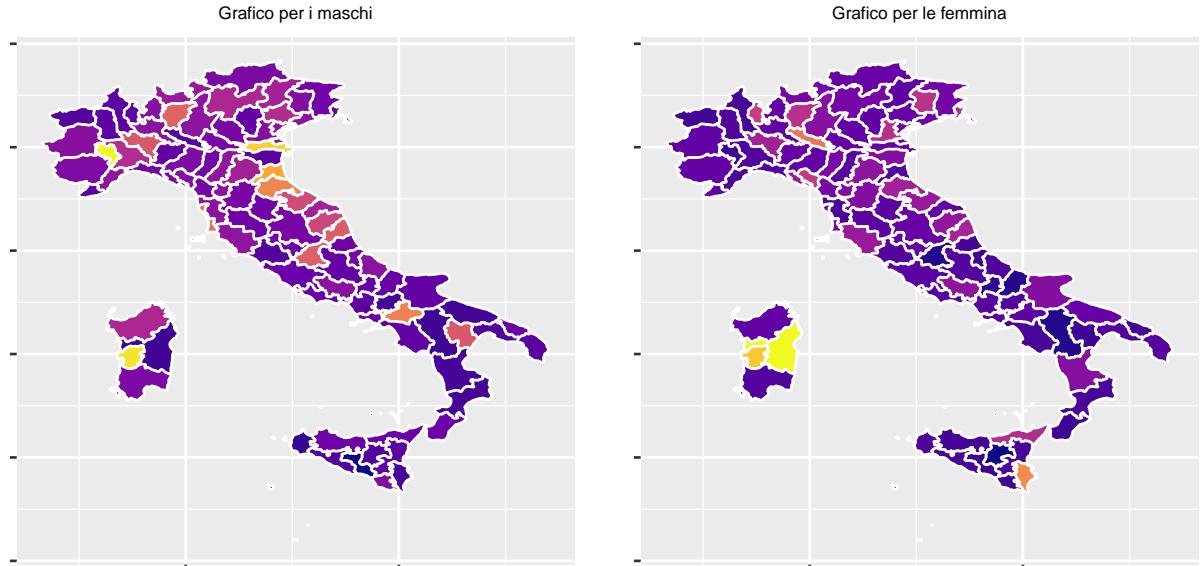
- Per la variabile **Capoluogo** sono le seguenti:



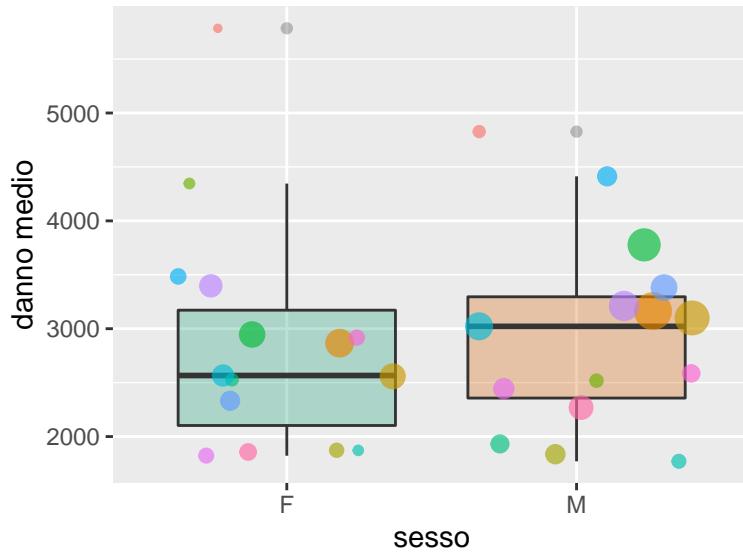
Mentre il boxplot con divisione per cluster provinciali evidenziati per numerosità è il seguente:



- Per la variabile **Sesso** sono le seguenti:

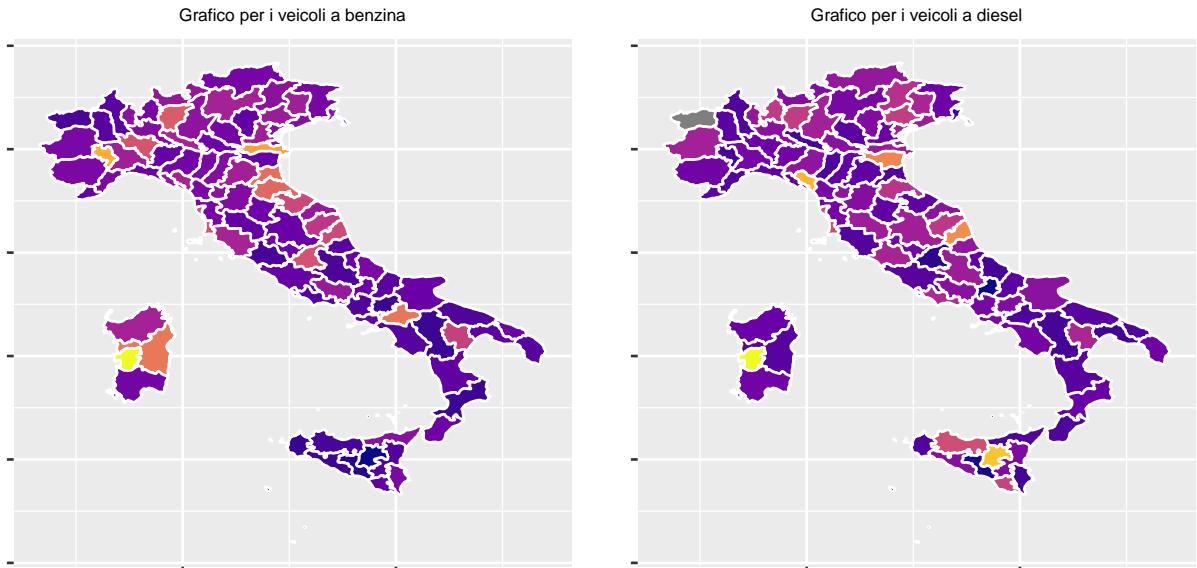


Mentre il boxplot con divisione per cluster provinciali evidenziati per numerosità è il seguente:

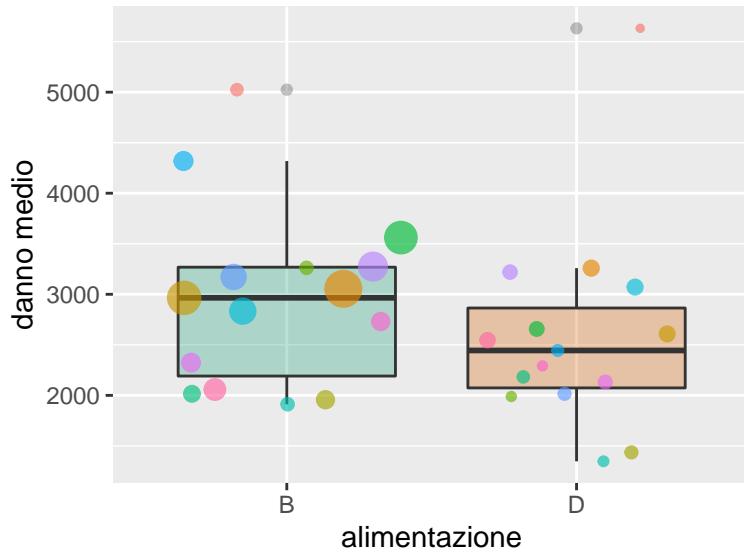


Per la variabile **Sesso** si ricorda che nella fase di pricing per ragioni normative sulla gender equality non può essere utilizzata.

- Per la variabile **Bendie** sono le seguenti:



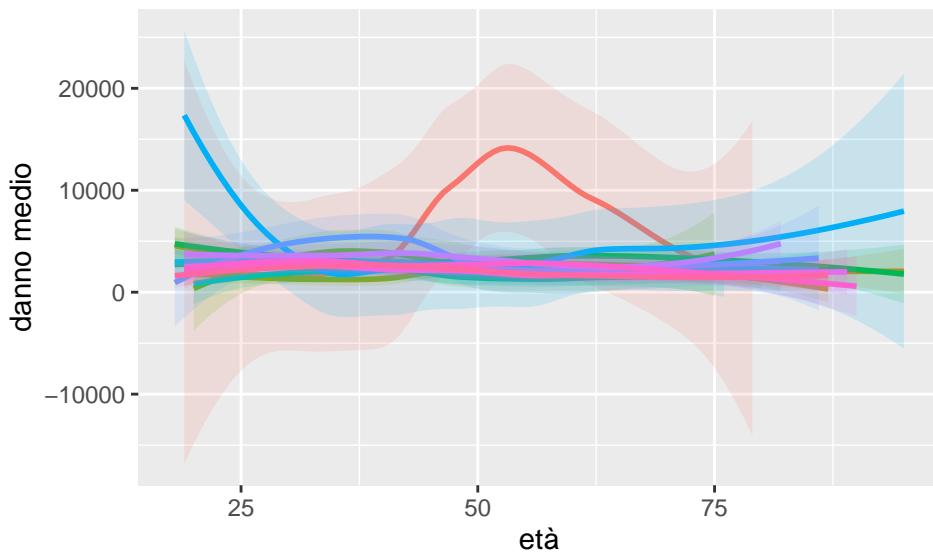
Mentre il boxplot con divisione per cluster provinciali evidenziati per numerosità è il seguente:



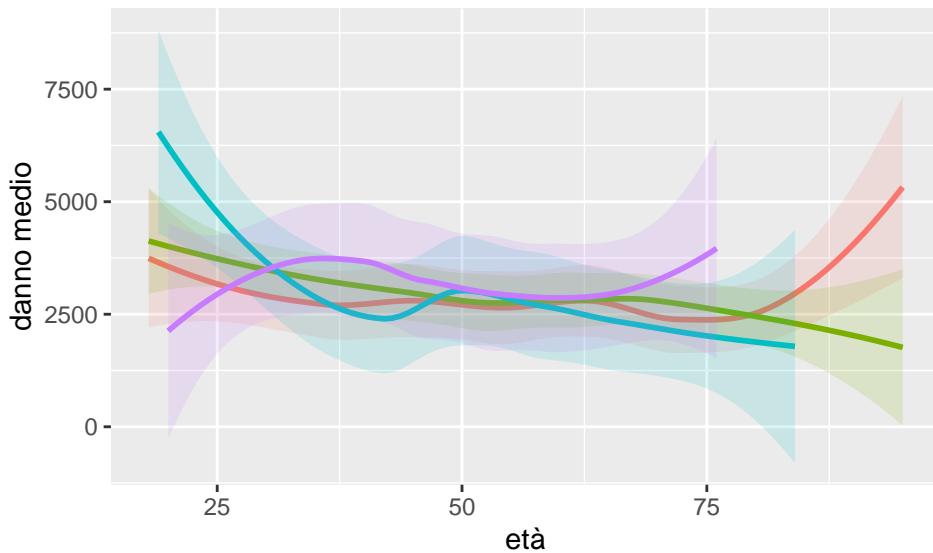
3.2. Relazione tra variabili esplicative

Si studiano le relazioni tra le diverse variabili esplicative. In particolare, si tiene conto della nuova fattorizzazione delle variabili **Eta**, **Prov** e **Potf**.

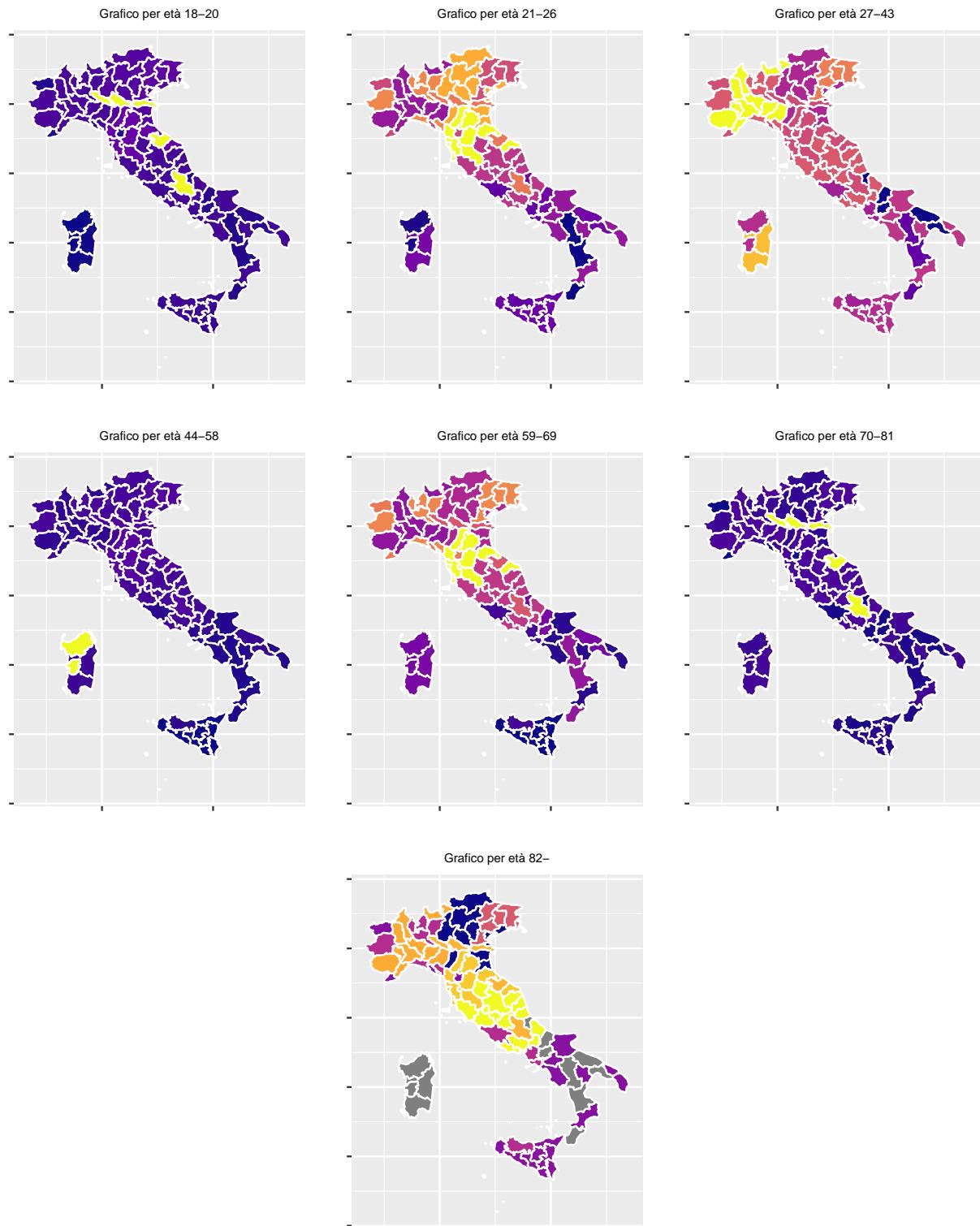
Il grafico per la variabile **Dannomedio** al variare dell'**Eta** per i diversi **cluster** provinciali mostra come il cluster provinciale 1 per età 45-60 abbia dei valori significativamente diversi rispetto gli altri clusters.



Analoghi grafici vengono creati per la variabile **Dannomedio** al variare dell'**Eta** per i diversi cluster della variabile **levepotf**.



Con gli stessi dati è possibile raccogliere le informazioni e rappresentarle per singoli clusters provinciali.



I dati possono essere rappresentati per **levelpotf** all'interno dei clusters provinciali.

Grafico per potenza fiscale -14

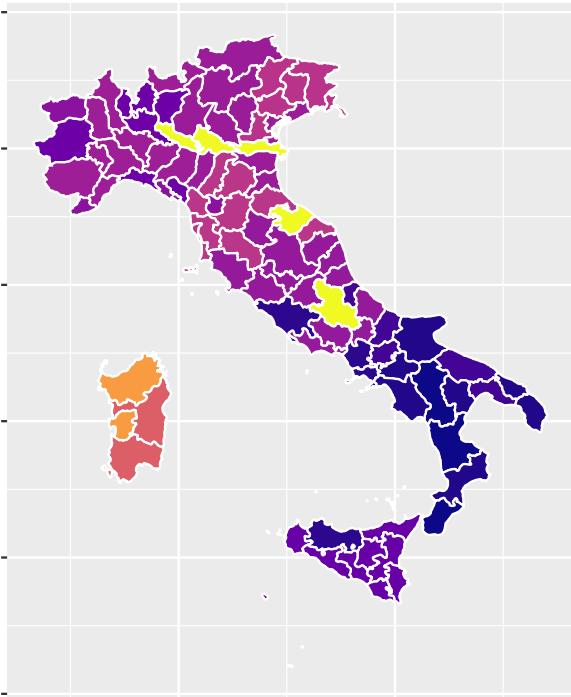


Grafico per potenza fiscale 15–18

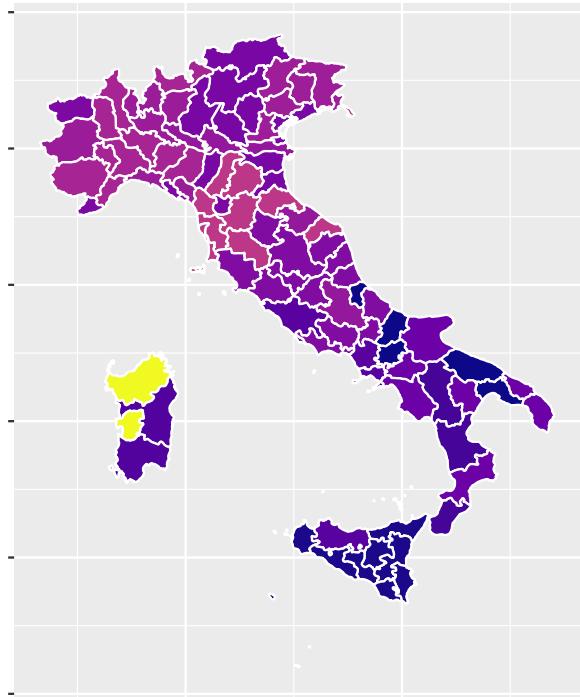


Grafico per potenza fiscale 19–21

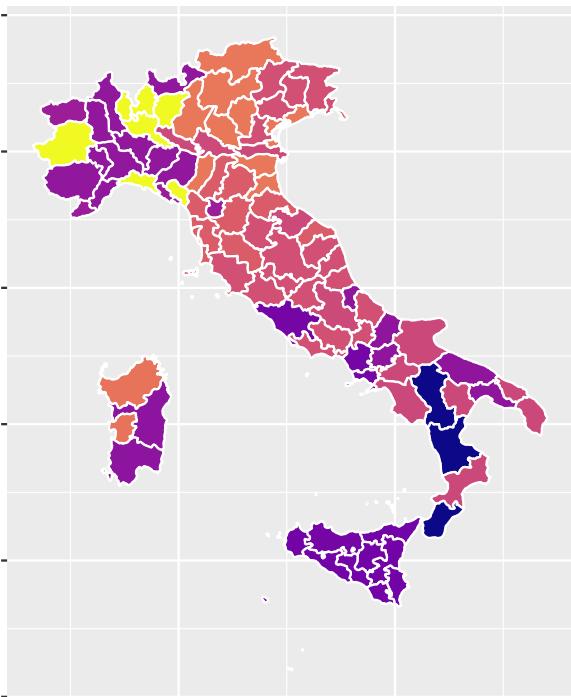
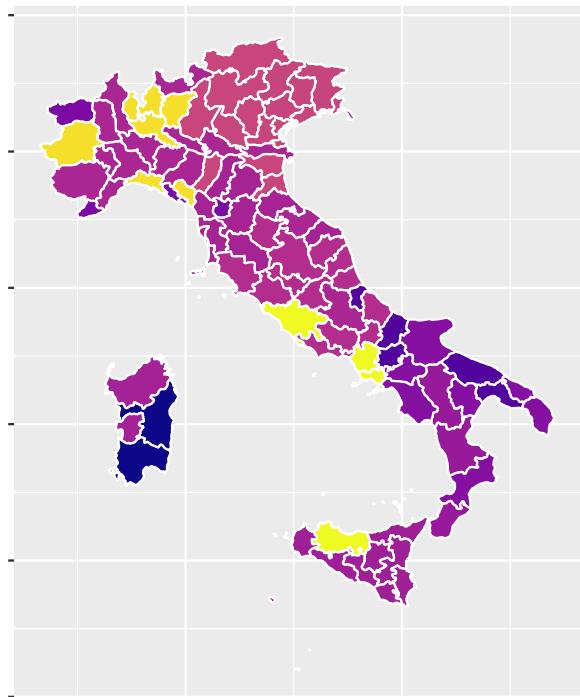
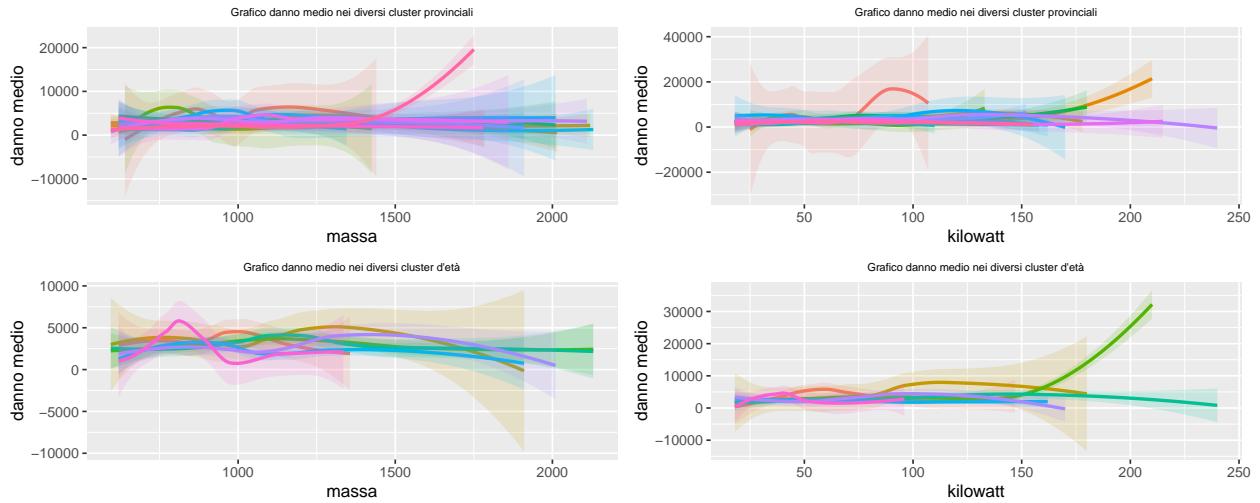


Grafico per potenza fiscale 22–



In conclusione, viene ripresa l'analisi delle variabili **Potf**, **Massa** e **Potkill**. Infatti, le variabili hanno un'alta correlazione: la correlazione **Massa** e **Potf** è 0.9024503 mentre per le variabili **Potkil** e **Potf** è uguale a 0.8442053. Le variabili **Massa** e **Potkill** non vengono prese in considerazione, anche per la difficoltà nel raggrupparle in clusters.



4. Modellizzazione con l'uso dei GLM

Per analizzare la variabile dipendente, si utilizzano i modelli lineari generalizzati. Questi modelli appartengono alla famiglia esponenziale, la cui densità ha forma $f(x; \theta) = h(x)c(\theta)\exp(w(\theta)t(x))$. Un GLM è caratterizzato da:

- la variabile dipendente deve appartenere alla famiglia esponenziale $Y_i \sim EF$;
- la variabile dipendente è affetta dalle covariate in modo lineare. La funzione, che prende il nome di preditore lineare, ha funzione $\eta_i = x_i^T \beta = \sum \beta_i x_i$;
- l'esistenza di una funzione, chiamata funzione collegamento, che connette la speranza matematica al preditore lineare $\mu_i = g(\eta_i)$ (quindi $\eta_i = g(\mu_i)^{-1}$).

Ora si mostrano le fasi del procedimento:

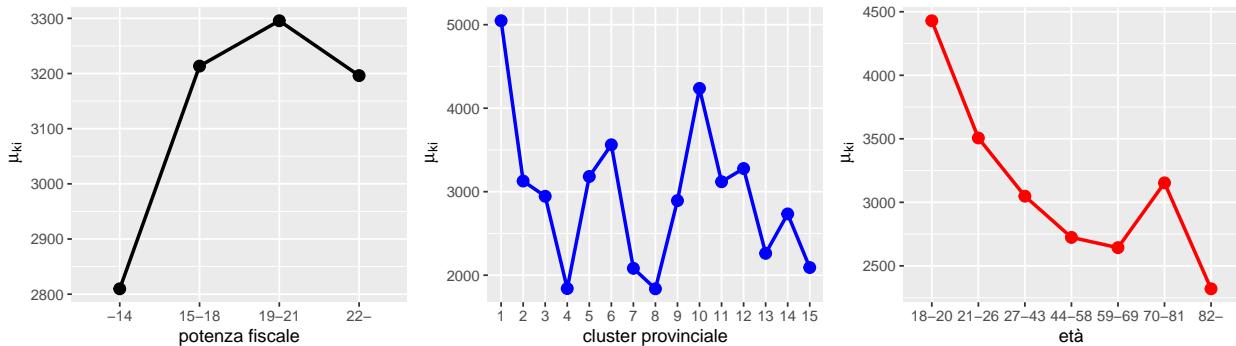
- modello Gamma con dati individuali;
- modello Gamma con dati raggruppati;

4.1. Modello Gamma con dati individuali

Il primo modello preso in considerazione è un GLM per la descrizione del danno per sinistro. Si considera il numero aleatorio Y_{ki} che rappresenta il danno aleatorio provocato dall'i-esimo sinistro in classe tariffaria k (con $k = 1,.., K$ e $i = 1,.., n_k$). Si ha dunque $f(y; \theta_k; \phi) = \exp\{\frac{1}{\phi}[y\theta_k + \log(-\theta_k)]c(y, \phi)\}$, dove la funzione cumulante della famiglia gamma è $b(\theta) = -\log(-\theta)$.

In particolare, si ha $E(Y_{ki}) = \mu_k$ e $V(Y_{ki}) = \phi V(\mu_k) = \phi \mu_k^2$. Non viene utilizzando il legame canonico, $g(\mu) = -\frac{1}{\mu}$, bensì il logaritmo come funzione di collegamento. Infatti, il legame canonico richiederebbe dei vincoli sui regressori affinché il previsore lineare sia positivo⁴.

Si provano alcuni modelli semplici con un'unica variabile per controllare l'effetto singolarmente delle variabili. Questa è la rappresentazione grafica delle principali variabili, cioè **levelpotf**, **cluster** e **leveleta**.



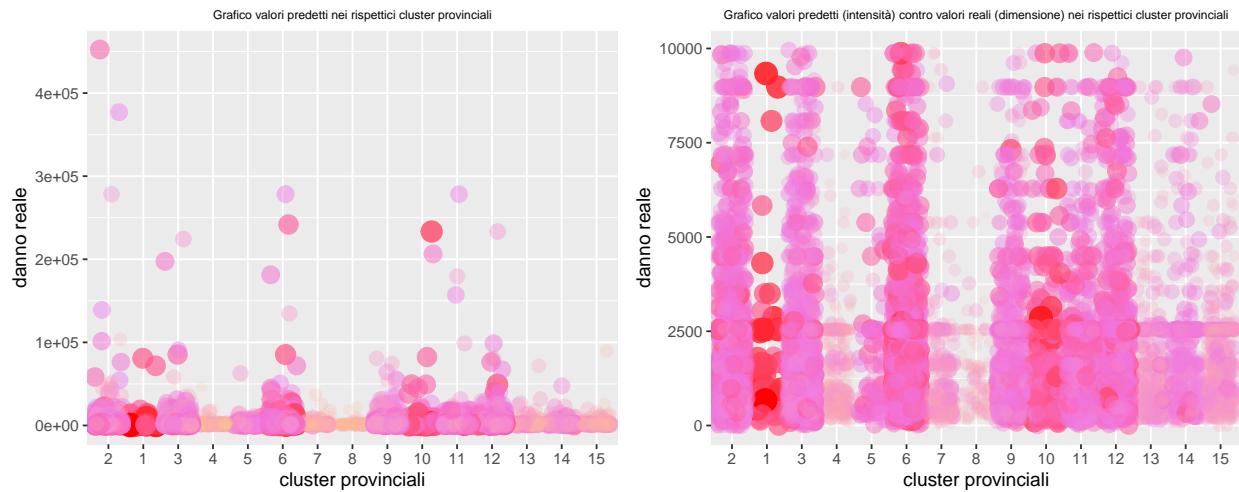
⁴ Si avrebbe che $E(Y_i) = g^{-1}(\eta_i) = \mu_i$, dove $\mu_i = -\frac{1}{\eta_i}$ può assumere valori negativi, Mentre la speranza matematica della distribuzione Gamma è sempre positiva.

Il modello stimato è il seguente:

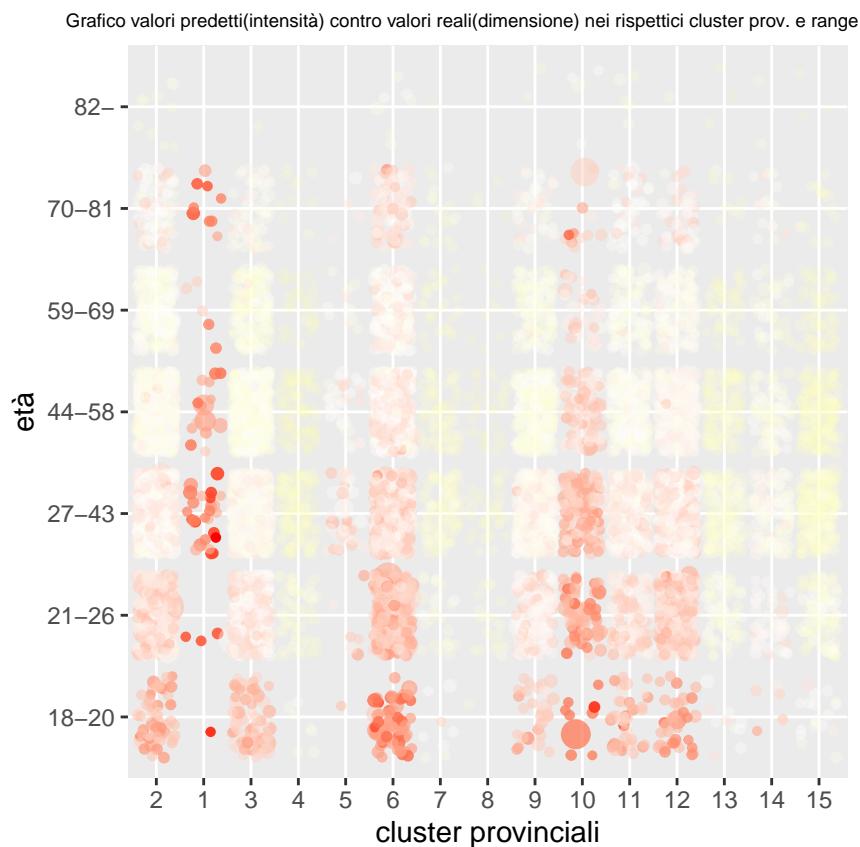
Criteri di valutazione della bontà di adattamento			
Criterio	DF	Valore	Valore/DF
Devianza	12E3	16155.3127	1.3782
Dev. scalata	12E3	13823.6164	1.1793
Chi-quadrato di Pearson	12E3	109389.9943	9.3320
X2 Pearson scal.	12E3	93601.7365	7.9851
Log verosimiglianza		-105513.4859	
Log verosimiglianza piena		-105513.4859	
AIC (minore è meglio)		211080.9717	
AICC (minore è meglio)		211081.1007	
BIC (minore è meglio)		211280.0006	

Analisi delle stime dei parametri di massima verosimiglianza							
Parametro		DF	Stima	Errore standard	Limiti di confidenza di Wald al 95%	Chi-quadrato di Wald	Pr > ChiQuadr
Intercept		1	8.5154	0.0629	8.3920 8.6387	18312.0	<.0001
leveleta	21-26	1	-0.2180	0.0623	-0.3401 -0.0959	12.25	0.0005
leveleta	27-43	1	-0.3535	0.0598	-0.4708 -0.2362	34.90	<.0001
leveleta	44-58	1	-0.4711	0.0606	-0.5897 -0.3524	60.50	<.0001
leveleta	59-69	1	-0.4976	0.0650	-0.6250 -0.3703	58.66	<.0001
leveleta	70-81	1	-0.3326	0.0766	-0.4828 -0.1825	18.85	<.0001
leveleta	82-	1	-0.6242	0.1825	-0.9818 -0.2666	11.70	0.0006
leveleta	18-20	0	0.0000	0.0000	0.0000 0.0000	.	.
Capoluogo	SI	1	-0.1421	0.0246	-0.1904 -0.0938	33.26	<.0001
Capoluogo	NO	0	0.0000	0.0000	0.0000 0.0000	.	.
Bendie	D	1	-0.2043	0.0374	-0.2776 -0.1311	29.90	<.0001
Bendie	B	0	0.0000	0.0000	0.0000 0.0000	.	.
Cluster	1	1	0.5897	0.1417	0.3121 0.8674	17.33	<.0001
Cluster	3	1	-0.0922	0.0335	-0.1577 -0.0266	7.59	0.0059
Cluster	4	1	-0.4789	0.0628	-0.6020 -0.3559	58.19	<.0001
Cluster	5	1	0.1495	0.1155	-0.0768 0.3758	1.68	0.1955
Cluster	6	1	0.1379	0.0343	0.0708 0.2051	16.21	<.0001
Cluster	7	1	-0.3356	0.0717	-0.4762 -0.1951	21.92	<.0001
Cluster	8	1	-0.5023	0.1094	-0.7167 -0.2879	21.08	<.0001
Cluster	9	1	-0.0495	0.0403	-0.1286 0.0296	1.51	0.2197
Cluster	10	1	0.2667	0.0617	0.1458 0.3876	18.70	<.0001
Cluster	11	1	-0.0164	0.0433	-0.1012 0.0684	0.14	0.7042
Cluster	12	1	0.0392	0.0376	-0.0345 0.1128	1.09	0.2972
Cluster	13	1	-0.3341	0.0587	-0.4492 -0.2190	32.38	<.0001
Cluster	14	1	-0.0796	0.0672	-0.2113 0.0522	1.40	0.2368
Cluster	15	1	-0.3358	0.0491	-0.4320 -0.2396	46.81	<.0001
Cluster	2	0	0.0000	0.0000	0.0000 0.0000	.	.
levelpotf	-14	1	-0.1686	0.0226	-0.2129 -0.1243	55.75	<.0001
levelpotf	19-21	1	0.0756	0.0336	0.0098 0.1414	5.08	0.0242
levelpotf	22-	1	0.2247	0.0673	0.0927 0.3566	11.14	0.0008
levelpotf	15-18	0	0.0000	0.0000	0.0000 0.0000	.	.
Scala		1	0.8557	0.0097	0.8369 0.8749	.	.

Alcune rappresentazioni grafiche, mostrano come i valori stimati si distribuiscano rispetto i diversi cluster provinciali in ascissa e il valore osservato in ordinata. Il colore e l'intensità dipendono dal valore della stima, più tendente al rosso e più intenso sono, maggiore è la loro stima.



Un ulteriore grafico mostra i valori stimati nei rispettivi cluster provinciali e di età, dove il colore dipende dal valore stimato mentre la dimensione dal valore reale osservato.



4.2. Modello Gamma con dati raggruppati

Il modello che si prende in considerazione è un GLM per i danni medi per sinistro nelle classi. Si procede al raggruppamento dei dati e si ha $Y_k = \frac{1}{m_k} \sum_{i=1}^{m_k} Y_{ki}$. La distribuzione è $\exp\{\frac{1}{m_k}[y\theta_k - b(\theta_k)]\}c(y, \phi, m_k)$. Inoltre, fissato il generico peso m_k per la classe k , si trova che $E(Y_k) = \mu_k$ e $V(Y_k) = \frac{\phi}{m_k}V(\mu_k)$. Anche in questo caso, si utilizza il collegamento logaritmico anzichè il canonico.

Le prime 5 osservazioni sono le seguenti:

Oss	Capoluogo	Bendie	Cluster	leveleta	levelpotf	nsin	dannocum	dannocummed
1	NO	B	1	21-26	-14	3	4358.95	1452.98
2	NO	B	1	27-43	-14	7	17996.84	2570.98
3	NO	B	1	27-43	15-18	7	19826.24	2832.32
4	NO	B	1	27-43	19-21	1	664.26	664.26
5	NO	B	1	44-58	-14	5	91726.54	18345.31

Il modello stimato è:

Criteri di valutazione della bontà di adattamento			
Criterio	DF	Valore	Valore/DF
Devianza	759	2192.5077	2.8887
Dev. scalata	759	893.5642	1.1773
Chi-quadrato di Pearson	759	3906.6508	5.1471
X2 Pearson scal.	759	1592.1692	2.0977
Log verosimiglianza		-6956.8893	
Log verosimiglianza piena		-6956.8893	
AIC (minore è meglio)		13967.7786	
AICC (minore è meglio)		13969.7759	
BIC (minore è meglio)		14093.7520	

Analisi delle stime dei parametri di massima verosimiglianza							
Parametro		DF	Stima	Errore standard	Limiti di confidenza di Wald al 95%	Chi-quadrato di Wald	Pr > ChiQuadr
Intercept		1	8.5154	0.0912	8.3366 8.6941	8721.95	<.0001
leveleta	21-26	1	-0.2180	0.0903	-0.3949 -0.0411	5.83	0.0157
leveleta	27-43	1	-0.3535	0.0867	-0.5234 -0.1835	16.62	<.0001
leveleta	44-58	1	-0.4711	0.0877	-0.6430 -0.2991	28.82	<.0001
leveleta	59-69	1	-0.4976	0.0941	-0.6821 -0.3131	27.94	<.0001
leveleta	70-81	1	-0.3326	0.1110	-0.5502 -0.1151	8.98	0.0027
leveleta	82-	1	-0.6242	0.2644	-1.1423 -0.1060	5.57	0.0182
leveleta	18-20	0	0.0000	0.0000	0.0000 0.0000	.	.
Capoluogo	SI	1	-0.1421	0.0357	-0.2121 -0.0721	15.84	<.0001
Capoluogo	NO	0	0.0000	0.0000	0.0000 0.0000	.	.
Bendie	D	1	-0.2043	0.0541	-0.3104 -0.0982	14.24	0.0002
Bendie	B	0	0.0000	0.0000	0.0000 0.0000	.	.

Cluster	1	1	0.5897	0.2053	0.1875	0.9920	8.26	0.0041
Cluster	3	1	-0.0922	0.0485	-0.1872	0.0029	3.61	0.0573
Cluster	4	1	-0.4789	0.0910	-0.6572	-0.3006	27.72	<.0001
Cluster	5	1	0.1495	0.1673	-0.1784	0.4774	0.80	0.3716
Cluster	6	1	0.1379	0.0496	0.0406	0.2352	7.72	0.0055
Cluster	7	1	-0.3356	0.1039	-0.5392	-0.1320	10.44	0.0012
Cluster	8	1	-0.5023	0.1585	-0.8129	-0.1916	10.04	0.0015
Cluster	9	1	-0.0495	0.0585	-0.1641	0.0651	0.72	0.3970
Cluster	10	1	0.2667	0.0894	0.0916	0.4419	8.91	0.0028
Cluster	11	1	-0.0164	0.0627	-0.1393	0.1065	0.07	0.7933
Cluster	12	1	0.0392	0.0544	-0.0675	0.1459	0.52	0.4718
Cluster	13	1	-0.3341	0.0851	-0.5008	-0.1673	15.42	<.0001
Cluster	14	1	-0.0796	0.0974	-0.2705	0.1114	0.67	0.4142
Cluster	15	1	-0.3358	0.0711	-0.4752	-0.1964	22.30	<.0001
Cluster	2	0	0.0000	0.0000	0.0000	0.0000	.	.
levelpotf	-14	1	-0.1686	0.0327	-0.2327	-0.1045	26.56	<.0001
levelpotf	19-21	1	0.0756	0.0486	-0.0197	0.1710	2.42	0.1199
levelpotf	22-	1	0.2247	0.0976	0.0335	0.4159	5.30	0.0213
levelpotf	15-18	0	0.0000	0.0000	0.0000	0.0000	.	.
Scala		1	0.4076	0.0184	0.3730	0.4453		

Alcuni livelli risultano non significativi. Si eseguono alcuni test per vedere se qualche acorpamento è possibile. In particolare, si testa singolarmente l'acorpamento per la variabile **cluster** dei livelli [5 6], [13 15], [2 9 11 12 14] e poi congiuntamente. Per la variabile **levelpotf** si raggruppano i livelli 19-21 e 22-. Successivamente si testano le modifiche delle due variabili congiuntamente. Il risultato è:

Risultati dei contrasti				
Contrasto	DF	Chi-quadrato	Pr > ChiQuadr	Tipo
cluster 5-6	1	0.00	0.9453	LR

Risultati dei contrasti				
Contrasto	DF	Chi-quadrato	Pr > ChiQuadr	Tipo
cluster 13-15 no 14	1	0.00	0.9864	LR

Risultati dei contrasti				
Contrasto	DF	Chi-quadrato	Pr > ChiQuadr	Tipo
cluster 2 9 11 12 14	4	2.56	0.6346	LR

Risultati dei contrasti				
Contrasto	DF	Chi-quadrato	Pr > ChiQuadr	Tipo
cluster 5-6 [2 9 11 12 14] 13-15	6	2.56	0.8614	LR

Risultati dei contrasti				
Contrasto	DF	Chi-quadrato	Pr > ChiQuadr	Tipo
potf 19-	1	2.18	0.1395	LR

Risultati dei contrasti				
Contrasto	DF	Chi-quadrato	Pr > ChiQuadr	Tipo
tutte	7	4.78	0.6871	LR

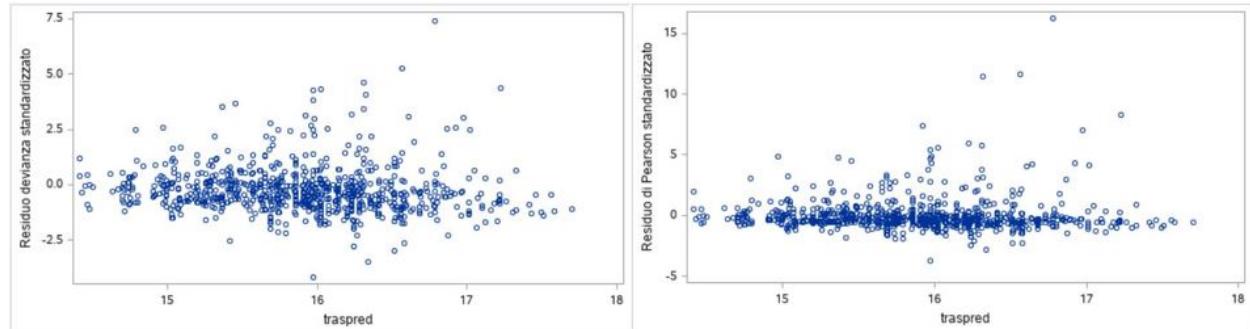
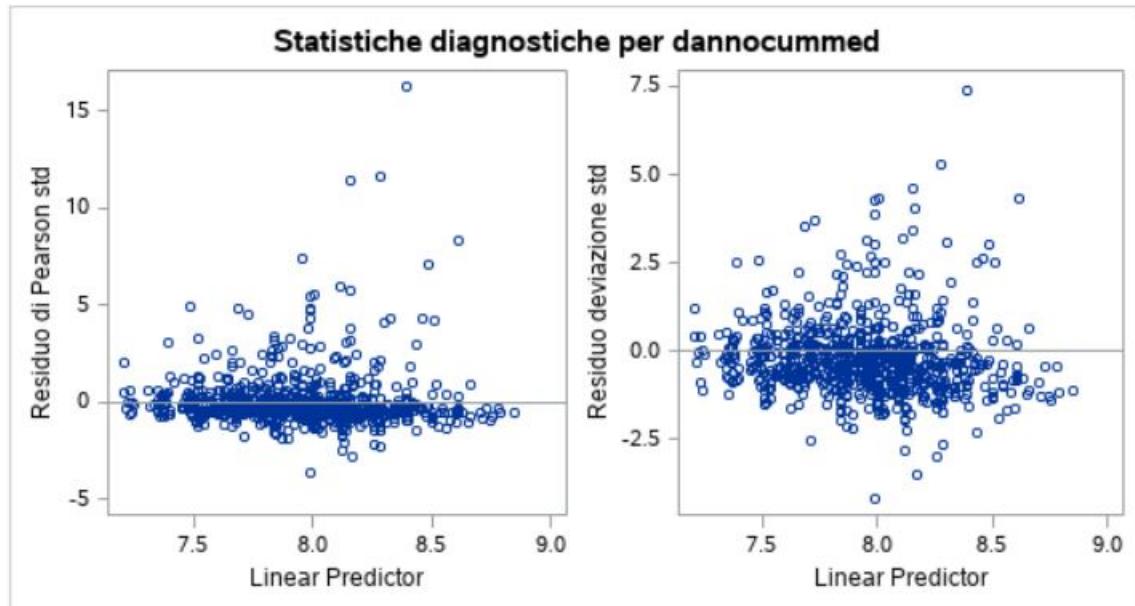
Con la nuova fattorizzazione, il modello stimato è:

Criteri di valutazione della bontà di adattamento				
Criterio	DF	Valore	Valore/DF	
Devianza	766	2204.2587	2.8776	
Dev. scalata	766	893.9726	1.1671	
Chi-quadrato di Pearson	766	3914.0288	5.1097	
X2 Pearson scal.	766	1587.3975	2.0723	
Log verosimiglianza		-6959.2780		
Log verosimiglianza piena		-6959.2780		
AIC (minore è meglio)		13958.5560		
AICC (minore è meglio)		13959.6555		
BIC (minore è meglio)		14051.8697		

Analisi delle stime dei parametri di massima verosimiglianza							
Parametro		DF	Stima	Errore standard	Limiti di confidenza di Wald al 95%	Chi-quadrato di Wald	Pr > ChiQuadr
Intercept		1	8.5068	0.0878	8.3348 8.6789	9394.85	<.0001
leveleta 21-26	1	-0.2187	0.0904	-0.3959	-0.0415	5.85	0.0156
leveleta 27-43	1	-0.3525	0.0869	-0.5227	-0.1823	16.47	<.0001
leveleta 44-58	1	-0.4697	0.0879	-0.6420	-0.2975	28.56	<.0001
leveleta 59-69	1	-0.4968	0.0943	-0.6816	-0.3120	27.75	<.0001
leveleta 70-81	1	-0.3323	0.1113	-0.5503	-0.1142	8.92	0.0028
leveleta 82-	1	-0.6288	0.2650	-1.1482	-0.1095	5.63	0.0176
leveleta 18-20	0	0.0000	0.0000	0.0000	0.0000	.	.
Capoluogo SI	1	-0.1430	0.0352	-0.2121	-0.0740	16.49	<.0001
Capoluogo NO	0	0.0000	0.0000	0.0000	0.0000	.	.
Bendie D	1	-0.1858	0.0518	-0.2873	-0.0843	12.88	0.0003
Bendie B	0	0.0000	0.0000	0.0000	0.0000	.	.

Cluster	1	1	0.5938	0.2043	0.1933	0.9943	8.45	0.0037
Cluster	3	1	-0.0870	0.0418	-0.1690	-0.0050	4.32	0.0376
Cluster	4	1	-0.4756	0.0878	-0.6476	-0.3036	29.36	<.0001
Cluster	5	1	0.1445	0.0423	0.0617	0.2273	11.69	0.0006
Cluster	6	1	-0.3291	0.1010	-0.5270	-0.1311	10.61	0.0011
Cluster	7	1	-0.4937	0.1569	-0.8012	-0.1862	9.90	0.0017
Cluster	8	1	0.2737	0.0859	0.1054	0.4420	10.16	0.0014
Cluster	9	1	-0.3277	0.0537	-0.4330	-0.2224	37.19	<.0001
Cluster	2	0	0.0000	0.0000	0.0000	0.0000	.	.
levelpotf	-14	1	-0.1677	0.0325	-0.2315	-0.1040	26.57	<.0001
levelpotf	19-	1	0.1024	0.0461	0.0121	0.1927	4.94	0.0263
levelpotf	15-18	0	0.0000	0.0000	0.0000	0.0000	.	.
Scala		1	0.4056	0.0183	0.3712	0.4432		

Si osservano i grafici dei residui di Pearson e di devianza rispetto al predittore lineare. Gli ultimi due grafici hanno in ascissa una trasformazione del predittore lineare, $2\log(\hat{\mu}_k)$, che permette una visualizzazione più facile. I punti del grafico si dispongono a “banda” senza mostrare curvature. Il modello è soddisfacente.



4.2.1. Modello Gamma con dati raggruppati e dispersione con stima di Pearson

Si visualizza la stima di un modello Gamma con dati raggruppati utilizzando lo stimatore di Pearson per la stima del parametro di dispersione ϕ . Infatti, dopo la stima del vettore $\hat{\beta}$, si procede alla stima della dispersione con lo stimatore $\tilde{\phi}_P = \frac{\tilde{X}^2}{n-p}$, dove $\tilde{X}^2 = \sum_{i=1}^n w_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$.

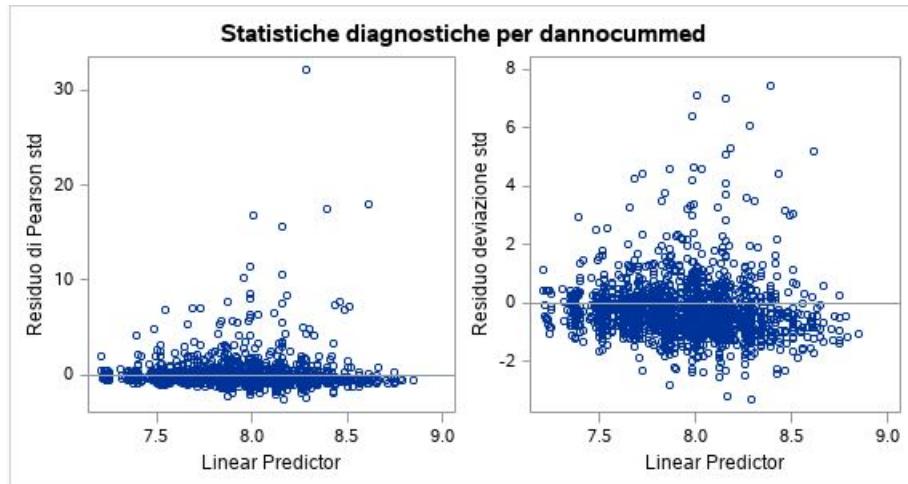
Il modello stimato è:

Criteri di valutazione della bontà di adattamento			
Criterio	DF	Valore	Valore/DF
Devianza	1784	4563.2955	2.5579
Dev. scalata	1784	1784.0000	1.0000
Chi-quadrato di Pearson	1784	12133.2698	6.8012
X2 Pearson scal.	1784	4743.4477	2.6589
Log verosimiglianza		-16144.1952	
Log verosimiglianza piena		-16144.1952	
AIC (minore è meglio)		32326.3903	
AICC (minore è meglio)		32326.8166	
BIC (minore è meglio)		32430.8373	

Analisi delle stime dei parametri di massima verosimiglianza							
Parametro		DF	Stima	Errore standard	Limiti di confidenza di Wald al 95%	Chi-quadrato di Wald	Pr > ChiQuadr
Intercept		1	8.5068	0.0894	8.3316 8.6820	9056.17	<.0001
leveleta	21-26	1	-0.2187	0.0921	-0.3992 -0.0382	5.64	0.0175
leveleta	27-43	1	-0.3525	0.0885	-0.5259 -0.1791	15.88	<.0001
leveleta	44-58	1	-0.4697	0.0895	-0.6452 -0.2943	27.53	<.0001
leveleta	59-69	1	-0.4968	0.0961	-0.6851 -0.3085	26.75	<.0001
leveleta	70-81	1	-0.3323	0.1133	-0.5544 -0.1101	8.60	0.0034
leveleta	82-	1	-0.6288	0.2699	-1.1578 -0.0999	5.43	0.0198
leveleta	18-20	0	0.0000	0.0000	0.0000 0.0000	.	.
Capoluogo	SI	1	-0.1430	0.0359	-0.2134 -0.0727	15.90	<.0001
Capoluogo	NO	0	0.0000	0.0000	0.0000 0.0000	.	.
Bendie	D	1	-0.1858	0.0527	-0.2892 -0.0825	12.42	0.0004
Bendie	B	0	0.0000	0.0000	0.0000 0.0000	.	.

Cluster	1	1	0.5938	0.2081	0.1859	1.0017	8.14	0.0043
Cluster	3	1	-0.0870	0.0426	-0.1705	-0.0035	4.17	0.0412
Cluster	4	1	-0.4756	0.0894	-0.6508	-0.3004	28.31	<.0001
Cluster	5	1	0.1445	0.0430	0.0601	0.2288	11.27	0.0008
Cluster	6	1	-0.3291	0.1029	-0.5307	-0.1274	10.23	0.0014
Cluster	7	1	-0.4937	0.1598	-0.8069	-0.1805	9.54	0.0020
Cluster	8	1	0.2737	0.0875	0.1023	0.4451	9.79	0.0018
Cluster	9	1	-0.3277	0.0547	-0.4350	-0.2204	35.85	<.0001
Cluster	2	0	0.0000	0.0000	0.0000	0.0000	.	.
levelpotf	-14	1	-0.1677	0.0331	-0.2327	-0.1028	25.61	<.0001
levelpotf	19-	1	0.1024	0.0469	0.0104	0.1944	4.76	0.0291
levelpotf	15-18	0	0.0000	0.0000	0.0000	0.0000	.	.
Scala		0	0.3909	0.0000	0.3909	0.3909		

Nota: The Gamma scale parameter was estimated by DOF/DEVIANCE.



4.2.2. Modello Gamma con dati raggruppati e dispersione con stima di quasi-devianza

Si procede alla stima del modello utilizzando lo stimatore di quasi-devianza. Per il parametro di dispersione si utilizza lo stimatore $\tilde{\phi}_D = \frac{\tilde{D}(c, f)}{n-p}$, dove $\tilde{D}(c, f) = -2 \sum_{i=1}^n w_i \{(y_i \hat{\theta}_i - b(\hat{\theta}_i)) - (y_i \theta_i^* - b(\theta_i^*))\}$, dove θ_i^* è il parametro canonico⁵ utilizzando il modello saturo con i tanti parametri quante le osservazioni.

Il modello stimato è:

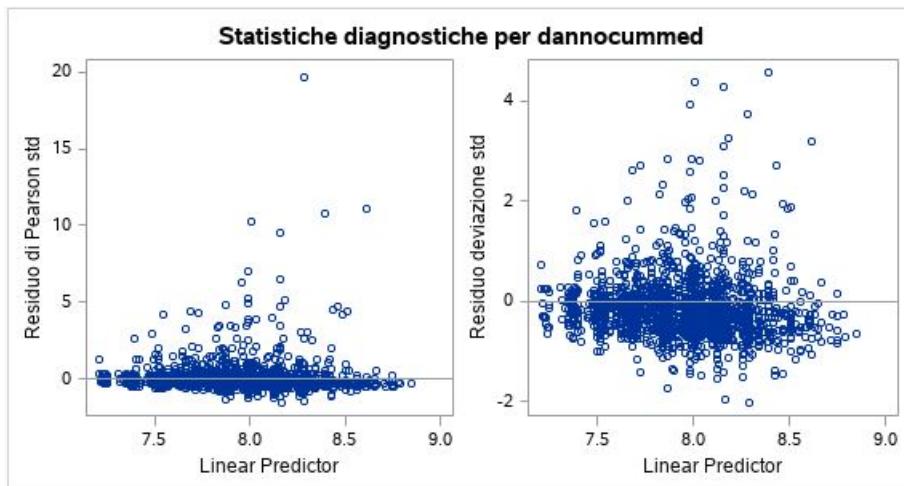
Criteri di valutazione della bontà di adattamento			
Criterio	DF	Valore	Valore/DF
Devianza	1784	4563.2955	2.5579
Dev. scalata	1784	670.9584	0.3761
Chi-quadrato di Pearson	1784	12133.2698	6.8012
X2 Pearson scal.	1784	1784.0000	1.0000
Log verosimiglianza		-16690.1689	
Log verosimiglianza piena		-16690.1689	
AIC (minore è meglio)		33418.3378	
AICC (minore è meglio)		33418.7641	
BIC (minore è meglio)		33522.7848	

Analisi delle stime dei parametri di massima verosimiglianza							
Parametro		DF	Stima	Errore standard	Limiti di confidenza di Wald al 95%	Chi-quadrato di Wald	Pr > ChiQuadr
Intercept		1	8.5068	0.1458	8.2211 8.7925	3406.01	<.0001
leveleta	21-26	1	-0.2187	0.1501	-0.5130 0.0756	2.12	0.1452
leveleta	27-43	1	-0.3525	0.1443	-0.6352 -0.0698	5.97	0.0145
leveleta	44-58	1	-0.4697	0.1460	-0.7559 -0.1836	10.35	0.0013
leveleta	59-69	1	-0.4968	0.1566	-0.8038 -0.1898	10.06	0.0015
leveleta	70-81	1	-0.3323	0.1848	-0.6945 0.0299	3.23	0.0722
leveleta	82-	1	-0.6288	0.4401	-1.4914 0.2337	2.04	0.1530
leveleta	18-20	0	0.0000	0.0000	0.0000 0.0000	.	.
Capoluogo	SI	1	-0.1430	0.0585	-0.2577 -0.0284	5.98	0.0145
Capoluogo	NO	0	0.0000	0.0000	0.0000 0.0000	.	.
Bendie	D	1	-0.1858	0.0860	-0.3543 -0.0173	4.67	0.0307
Bendie	B	0	0.0000	0.0000	0.0000 0.0000	.	.

⁵ Non avendo utilizzato il collegamento canonico non vale la relazione $\phi_i = b'^{-1}(g^{-1}(x'_i \beta)) = x'_i \beta = \eta_i$.

Cluster	1	1	0.5938	0.3393	-0.0713	1.2589	3.06	0.0801
Cluster	3	1	-0.0870	0.0695	-0.2232	0.0492	1.57	0.2107
Cluster	4	1	-0.4756	0.1458	-0.7613	-0.1899	10.65	0.0011
Cluster	5	1	0.1445	0.0702	0.0069	0.2820	4.24	0.0395
Cluster	6	1	-0.3291	0.1678	-0.6579	-0.0003	3.85	0.0498
Cluster	7	1	-0.4937	0.2606	-1.0044	0.0170	3.59	0.0581
Cluster	8	1	0.2737	0.1426	-0.0058	0.5532	3.68	0.0550
Cluster	9	1	-0.3277	0.0892	-0.5026	-0.1528	13.48	0.0002
Cluster	2	0	0.0000	0.0000	0.0000	0.0000	.	.
levelpotf	-14	1	-0.1677	0.0540	-0.2737	-0.0618	9.63	0.0019
levelpotf	19-	1	0.1024	0.0765	-0.0476	0.2524	1.79	0.1809
levelpotf	15-18	0	0.0000	0.0000	0.0000	0.0000	.	.
Scala		0	0.1470	0.0000	0.1470	0.1470		

Nota: The Gamma scale parameter was estimated by DOF/Pearson's Chi-Square



5. Conclusioni

In conclusioni, i modelli con dati raggruppati si stetizzano in maniera migliore il problema. Tra i tre modelli Gamma con dati raggruppati - avendo le stesse stime, come era atteso- si hanno per ogni parametro, un valore del p-value differente. Questa differenza è dovuta al fatto che si utilizzano diverse stime del parametro di dispersione. In questo caso, il primo modello, 4.1. , risulta il più ottimale: sia per il fatto che i parametri sono tutti significativi, sia per il confronto del AIC, sia confrontando i dati sui residui.