Arrests 2010 NTA: Analisi

Analisi NTA 2010 - 2011 interazione tra variabili

Descrizione

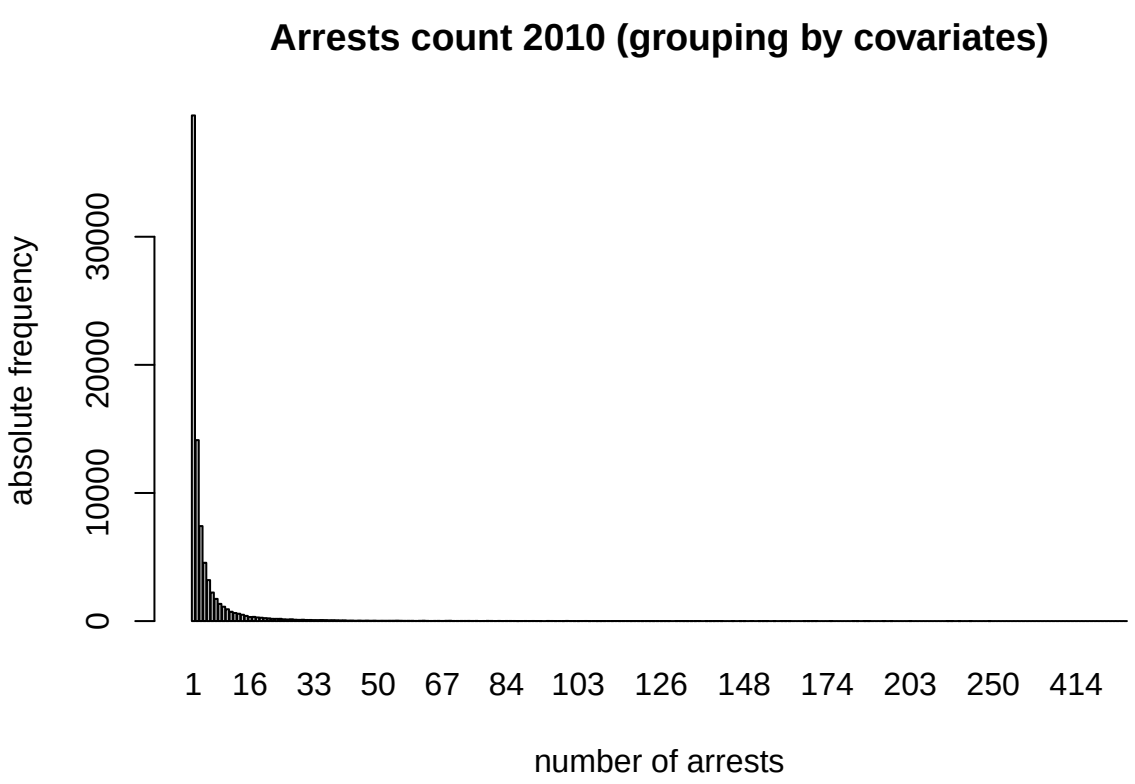
L'obbiettivo di questa sezione di analisi è verificare se esiste un sottoinsieme di variabili esplicative particolarmente correlate con il numero di arresti, sia marginalmente che considerando l'interazione con ciascuna zona spaziale (NTA). Per vincoli computazionali si riduce l'insieme di stima al solo anno 2010: per quest'anno i dati del censo sono esatti e non si sono verificati eventi rari a differenza del 2020 (Covid); l'insieme di verifica scelto è l'anno 2011, in quanto è l'anno più vicino al 2010 (l'assunzione è che i due anni siano abbastanza simili per il fenomeno considerato).

Problematiche

Questi dati presentano diverse problematiche.

Le scelte fatte sono dovute a fattori computazionali, di tempo e al fatto che per permettere conteggi diversi da 1 è necessario considerare zone spaziali e intervalli temporali non eccessivamente ristretti.

Per quanto concerne gli intervalli temporali si è scelto di ignorare i possibili trend e considerare i singoli anni, per ciascun anno si sono utilizzati i mesi per la costruzione degli insiemi di convalida incrociata. Pur avendo a disposizione il giorno di ciascun arresto la selezione dei mesi è apparsa come un giusto compromesso per garantire che non tutti i conteggi fossero uguali a 1 che comunque è il conteggio minino e più frequente molto superiore a tutti gli altri conteggi:



Questa "sovradispersione" di 1 è ragionevole data la costruzione dei conteggi per aggregazione di osservazioni con le stesse combinazioni di covariate; si dovrebbero inoltre aggiungere osservazioni con conteggi nulli per ogni combinazione di variabili per cui non si sono osservati arresti.

Mantenere tutti i conteggi unitari rende computazionalmente molto oneroso l'addattamenteo dei modelli e può creare problemi nella selezione degli stessi

La soluzione adottata è basata sul sottocampionamento: si stabilisce una soglia per i valori di conteggi oltre cui non sottocampionare, si conta la frequenza di conteggi osservati per tale soglia e si sottocampiona un sottoinsieme di grandezza uguale a quella frequenza da ciascun sottoinsieme di conteggi con valori inferiori alla soglia.

L'assunzione di fondo è che, almeno per i conteggi fino alla soglia considerata, la frequenza sia descrescente rispetto al valore degli stessi; un aspetto da sottolineare della metodologia proposta, in quanto compromesso, è che introduce distorsione nelle stime.

Si considerano due tipologie di dataset, entrambi impiegano il sottocampiomamento, ma in uno i conteggi nulli sono presenti e nell'altro sono assenti (i modelli per risposta continua sono adattati impiegando una trasformazione logaritmica dei dati senza conteggi nulli).

Per questo studio la soglia selezionata che è apparsa ragionevole in base alle considerazioni precedenti è di conteggi uguali a 10.

Il sottocampionamento è effettuato anche per i dataset completi relativi a 2010 e 2011 (escludendo i mesi come variabile di raggruppamento), ma non sono stati aggiunti gli zeri per permettere il confronto tra modelli per risposta continua.

Elevata dimensionalità

I dati presentano elevata dimensionalità considerando le interazioni tra la variabile spaziale (NTA) e le covariate (qualitative) di arrests. E' comunque interessante provare i metodi di selezione delle variabili anche sui dati senza interazioni.

Per avere delle misure quantitative si considera il dataset in cui si sono definiti i conteggi senza considerare i mesi: si riporta il rapporto tra il numero di osservazioni (righe) e il prodotto tra il numero di modalità di NTA e la somma delle modalità delle variabili qualitative di arrests.

Senza considerare interazioni tra KY_CD (esplicativa non spaziale di arrest con più modalità) con gli NTA il rapporto è (considerando i dati 2011 :

```
##  NTA2020
## 3.816125
```

Considerando anche interazioni tra KY_CD e NTA rapporto è:

```
##   NTA2020
## 0.9178023
```

Modelli

Criterio di selezione dei parametri di regolazione

Come già accennato, i parametri di regolazione sono selezionati tramite convalida incrociata (CV) impiegando i mesi per la costruzione degli insiemi.

La procedura per la costruzione degli insiemi è la seguente: - Selezione di k: il numero di insiemi di convalida (ad esempio k = 4) - Ogni insieme di convalida è composto da osservazioni raggruppate di 12 / k (3) mesi e i mesi rimanenti (9) vengono utilizzati per adattare il modello. - Per cercare di compensare e mediare le fluttuazioni stagionali, i mesi di validazione sono scelti il più distanziati possibile. Ad esempio, nel caso di k = 4, il primo insieme di validazione è (gennaio, maggio, settembre), il secondo set è (febbraio, giugno, ottobre), il terzo è (marzo, luglio, novembre) e il quarto è (aprile, agosto, dicembre). - Per rendere ogni risposta comparabile avendo utilizzato un numero diverso di mesi, una nuova risposta è definita come il rapporto degli arresti diviso per il numero di mesi utilizzati nel raggruppamento (ovvero l'esponenziale dell 'offset nel modello di Poisson).

Matrice del modello

La matrice del modello considerata è quella con tutte le variabili e le interazioni tra tutte le variabili di arrests tranne KY_CD (per ragioni computazionali) e le zone spaziali degli NTA.

Poichè la matrice del modello dell'insieme di verifica e quella dell'insieme di stima non condividono tutte le colonne si considerare solo le colonne in comune alle due.

```
## [1] 17039  4094
```

```
## [1] 17164  4109
```

Esplicative quantitative

IL dtaset presenta principalmente esplicative categoriali, benchè le due esplicative quantitative (MdAge) permettano la specificazione di diverse forme funzionali qui, per ragioni computazionali ci si limita ad assumere una relazione monotona lineare con la risposta.

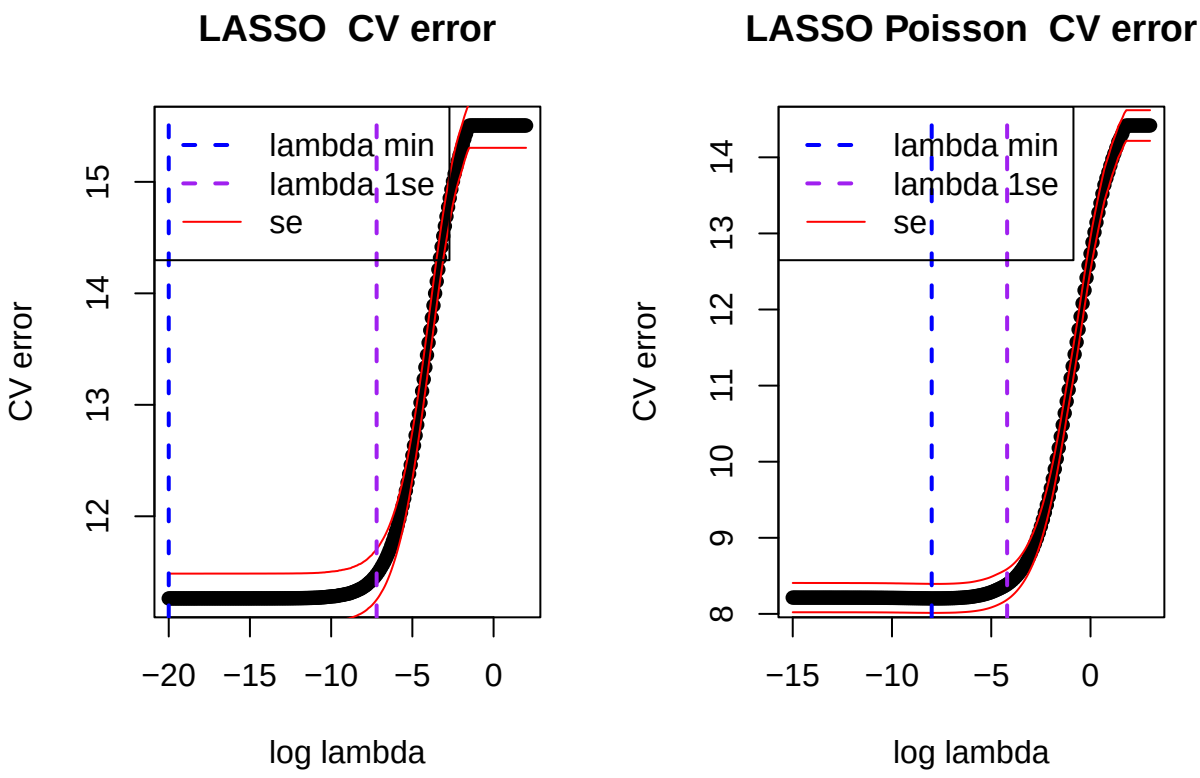
Modelli per risposta continua

Nell'impiego dei modelli con risposta continua (con errori gaussiani i.i.d) si è scelto di stimare il modello su una trasformazione logaritmica della risposta: "y = count / n_month_train" (per i dati senza introduzione di conteggi nulli) e calcolare l'errore di previsioni sulla trasformazione "count = exp(y) * n_month_test" rispetto al numero di conteggi osservati, in questo modo la previsione è sempre positiva.

Modelli e procedure considerati

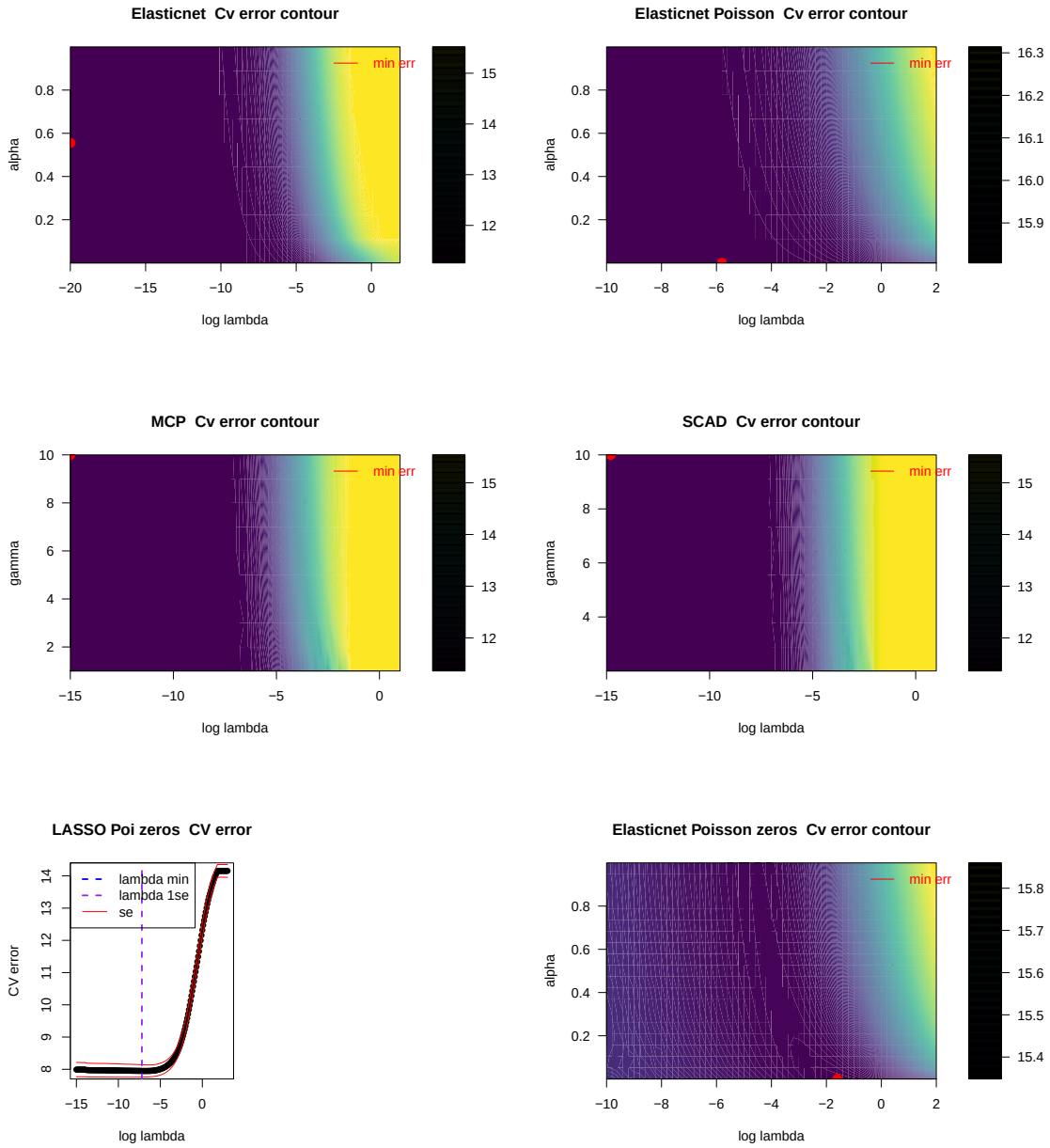
I modelli considerati sono modelli normali con penalizzazioni LASSO, Elasticnet, SCAD ed MCP e modelli Poisson con penalizzazioni LASSO ed Elasticnet. Per per tutti i modelli si seleziona il parametro (eventualmente vettoriale) di regolazione che minimizza l'errore di convalida. Per i metodi per cui il parametro di regolarizzazione ha dimensione 2 si definisce una griglia di valori (di cui si riporta il grafico delle curve di livello dell'errore). Per i metodi SCAD e MCP, poichè "ncvreg" presenta dei problemi computazionali dovute alle dimensioni del dataset è impiegata la libreria "picasso" che però non fornisce indicazioni rispetto alle regioni non convesse.

Per il modello normale il λ minimo è molto vicino a zero (poichè la soluzione è sul bordo si dovrebbe provare a diminuire ulteriormente λ, ma già così i coefficienti sono quasi uguali alle stime non penalizzate). Per il modello Poisson.



Elasticnet individua in entrambi i casi un α prossimo a zero (ridge, possibilmente data la forte correlazione tra covariate nella costruzione della matrice del modello), nel modello normale presenta una soluzione molto vicina alle stime non penalizzate (λ è prossimo a zero), mentre nel modello Poisson il λ selezionato è

```
## [1] 0.003027555
```



Per confontare modelli per risposta continua e discreta nelle previsioni sui dati 2011 si approssimano le previsioni continue al primo intero

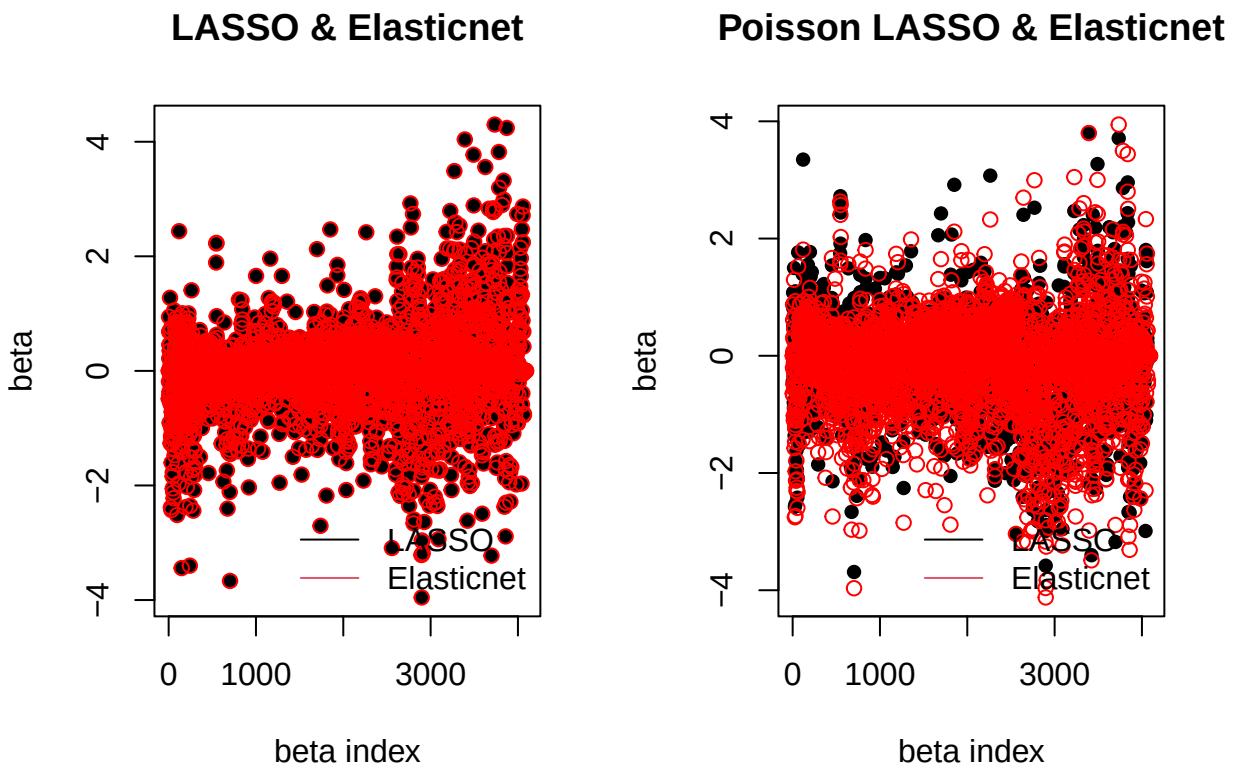
```
## Warning in rm(temp.fit): oggetto 'temp.fit' non trovato
```

Modelli migliori

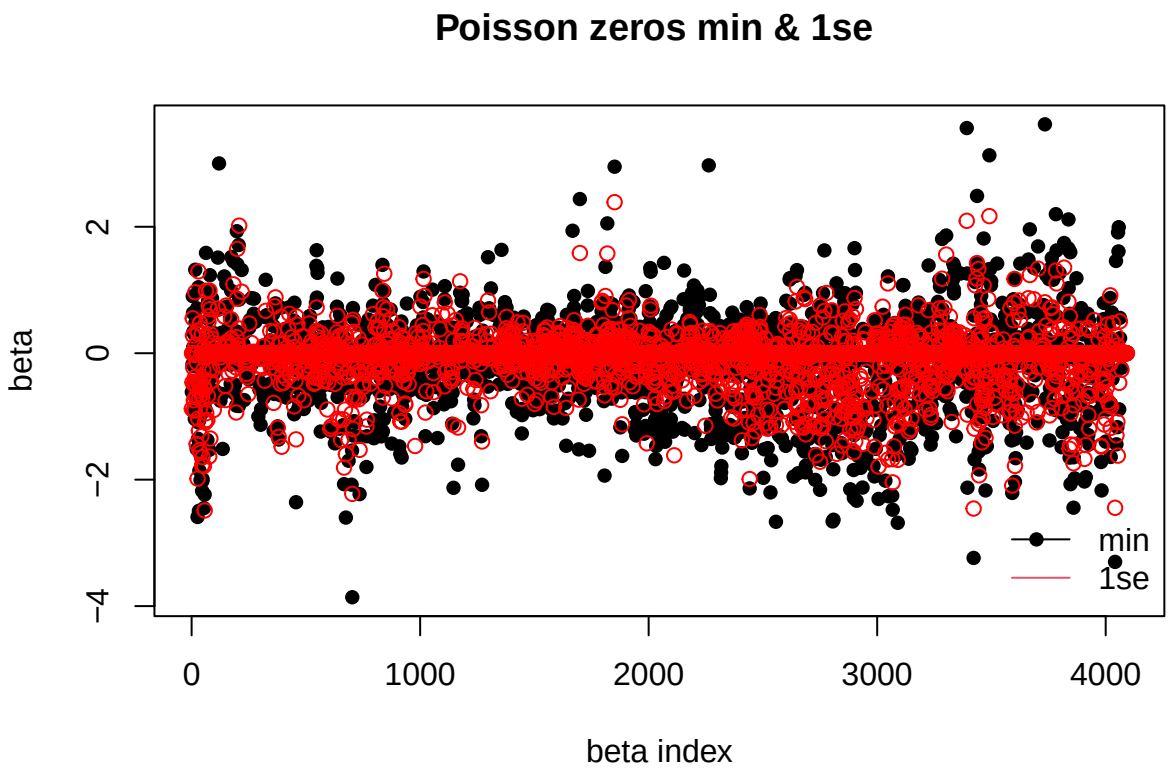
Si riportano gli errori di previsione sui dati del 2011 (senza zeri) dei vari modelli migliori stimati sui dati completi 2010 (senza zeri). La milgior previsione si ha per il modello Poisson con penalità LASSO (selezionato sui dati con gli zeri) per λ a errore a un errore standard, mentre il peggiore è sempre il modello di Poisson ma con penalità Elasticnet.

```
##                        model test_error
## 1                      lasso   21.27275
## 2                  lasso.1se   21.32034
## 3                 elasticnet   21.27275
## 4                       scad   20.24735
## 5                        mcp   20.24735
## 6              poisson_lasso   17.07644
## 7           poisson_lasso1se   30.24875
## 8         poisson_elasticnet   30.25221
## 9        poisson_lasso.zeros   16.97896
## 10  poisson_lasso.zeros.1se   16.76682
## 11 poisson_elasticnet_zeros   30.24320
```

I grafici dei coefficienti stimati confermano, per LASSO ed Elasticnet non avviene selezione delle variabili. Le stime non sono sparse nemmeno con il criterio dell'errore a un errore standard. Anche per SCAD ed MCP non avviene selezione di variabili: le stime (non riportate) sono quasi uguali a quelle LASSO ed Elasticnet.



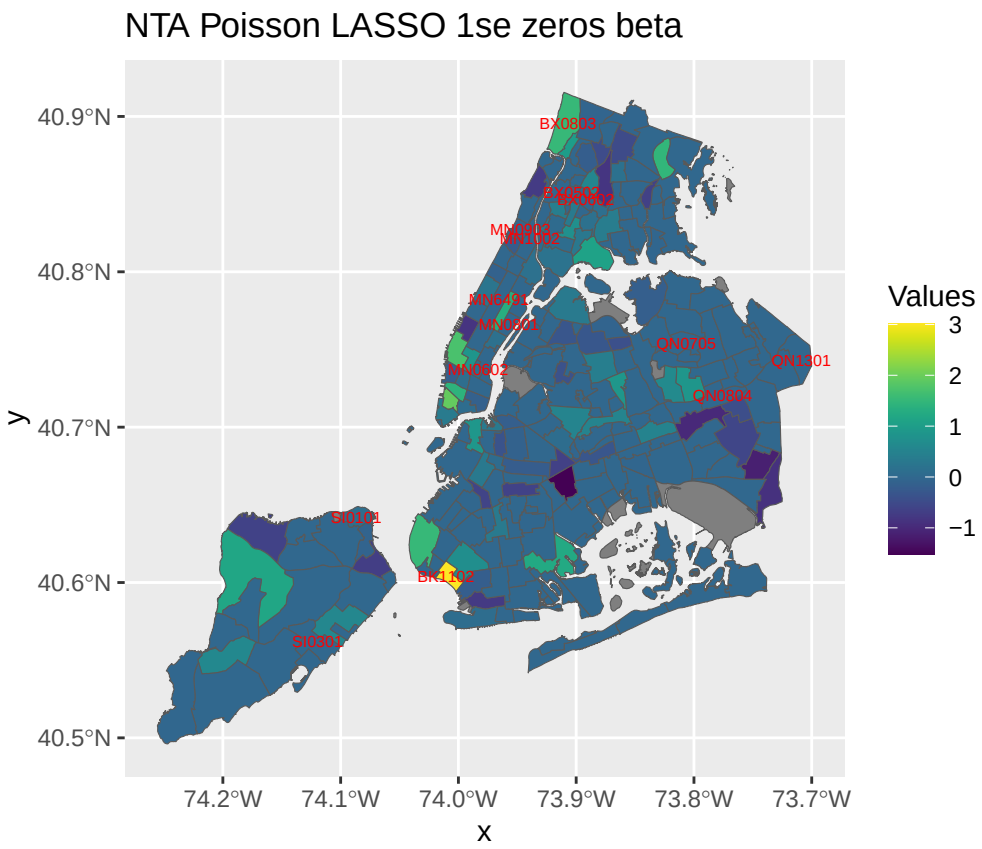
I modelli Poisson adattati considerando gli zeri presentano una maggiore selezione di variabili rispetto ai precedenti.



La tabella sottostante contiene per ciascun modello il rapporto tra il numero di elementi non nulli e il numero di elementi totali del vettore dei coefficienti: il modello Poisson con penalizzazione LASSO (criterio a un errore standard) e impiegando gli zeri è quello con il rapporto minore, questo modello è dunque l'oggetto delle successive analisi.

```
##                        model not_null_ratio
## 1                      lasso      0.7082824
## 2                  lasso.1se      0.5223552
## 3                 elasticnet      0.7082824
## 4                       scad      0.7082824
## 5                        mcp      0.7082824
## 6              poisson_lasso      0.6017591
## 7           poisson_lasso1se      0.4348888
## 8         poisson_elasticnet      0.7082824
## 9        poisson_lasso.zeros      0.5668214
## 10  poisson_lasso.zeros.1se      0.4273149
## 11 poisson_elasticnet_zeros      0.7082824
```

E' riportata la mappa degli NTA colorati in base al valore del corrispettivo coefficiente (marginale) (le zone grigie corrispondono a zone non presenti nei dati, e quindi coefficienti non stimati),per alcuni NTA sono presenti i nomi dei corrispettivi codici, questi sono in corrispondenza dei coefficienti più grandi relativi a termini di interazione (vedasi sotto).



Si riportano anche le tabelle dei cofficienti per le variabili qualitative (marginali).

Poichè si è più interessati ai termini correlati positivamente con il numero di arresti si riportano di seguito i 20 coefficienti più grandi del modello Poisson più sparso descritto sopra.

LASSO Poisson zeros 1