Incendi nel Montesinho Natural Park

Federico Boiocchi
Andrea Zanzottera

January 27, 2025

Abstract

Il Parco Naturale di Montesinho è un'area protetta di 75'000 ettari situata a Nord-Est del Portogallo. In questo report vogliamo analizzare, tramite una regressione di misture, la relazione tra l'intensità degli incendi registrati in questo parco e alcune variabili metereologiche ad essi associate. Abbiamo dapprima stimato le etichette con un modello MoE e successivamente le abbiamo utilizzate come ground-truth per fare classification. La nostra analisi si basa sul dataset **Forest Fires** scaricabile da *UC Irvine Machine Learning Depository.* Abbiamo infine avanzato diverse ipotesi in merito alla caratterizzazione delle classi non ottenendo però risultati soddisfacenti. Le informazioni relative alle variabili e ai loro dettagli sono basate sull'articolo *A Data Mining Approach to Predict Forest Fires using Meteorological Data* di P. Cortes e A. Morais

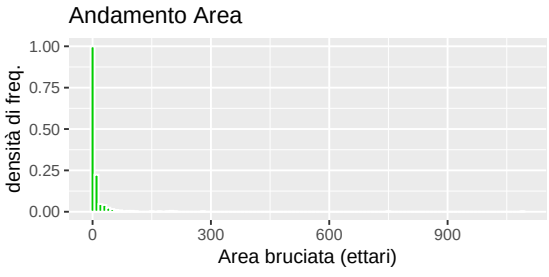
1 Forest Fires Dataset

Il dataset **Forest Fires** è costituito da 517 osservazioni e 13 variabili, è in forma tidy e non contiene missing values. Su ciascuna osservazione, rappresentante un incendio, sono stati misurati:

- 4 indici che abbiamo scartato per problemi di multicollinearità con le variabili misurate sperimentalmente.
- 3 variabili continue: **Temp** temperatura in °C, **Wind** velocità del vento in Km/h, e **RH** umidità relativa in % , interpretabili come regressori
- la variabile **Rain** scartata perchè zero-inflated
- la variabile **Area**, ossia la superficie in ettari bruciata dall'incendio, interpretabile come risposta
- le variabili **Day** e **Month**, mese e giorno dell'incendio
- le variabili integer **X** e **Y** rappresentanti la posizione dell'incendio sulla mappa del Montesihno Park rispetto a una griglia sovraimposta 9 x 9

2 Analisi Esplorativa

2.1 la variabile Area



Un primo problema riguarda la variabile risposta **Area**, la quale è fortemente asimmetrica e zero-inflated, come si può osservare dal grafico a sinistra. Inoltre non sembra derivare da una mistura di normali. Proponiamo una trasformazione $ln(Area + 1)$ per attenuare la asimmetria e permettere la stima del modello di regressione. Inoltre ipotizziamo che la alta presenza di zeri sia dovuta alla scarsa sensibilità della misurazione delle aree, in altre parole, roghi che hanno bruciato meno di $900m^2$ hanno un'area registrata nulla. Volendo avremmo potuto sostituire gli zeri con stime non nulle secondo distribuzioni suggerite da esperti. Non lo abbiamo fatto.

Nota: Osserviamo che una trasformata logaritmica può attenuare tale asimmetria. Al contempo, applicare solo $ln(Area)$, per via dei numerosi zeri, porterebbe a valori -Inf, (assolutamente non gestibili dagli algoritmi iterativi EM/IRLS usati per la stima dei modelli). Pertanto applichiamo una trasformazione $ln(Area + 1)$ in modo da attenuare l'asimmetria e permettere la stima. Abbiamo poi stimato la densità non-parametrica della variabile **log_area** la quale sembra derivare da una mistura di normali univariate a 3 componenti tagliata sullo 0. Si noti che la variabile trasformata rimane zero-inflated e pertanto la nostra intera analisi ne risentirà; sarebbe interessante ripetere l'analisi con modelli che tengono conto degli zeri (non utilizzati in questo report).

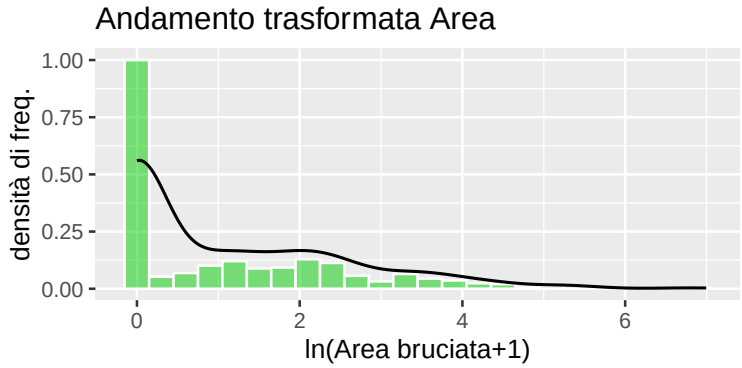


Figure 1: istogramma e densità non-parametrica (kernel gaussiano)

2.2 Le variabili esplicative

Tra le variabili esplicative contenute in **Forest Fires**, abbiamo scelto quelle ad elevata interpretabilità. E' utile osservare che abbiamo deciso di applicare una tecnica unsupervised (MoE e poi classification con etichette stimate) nonostante avessimo le etichette vere (**Day**, e **Month**) dal momento che sarebbero venuti troppi cluster (7 o 12) difficilmente stimabili con sole 517 osservazioni. Inoltre, considerando l'etichetta più interpretabile, ossia **Month**, è plausibile pensare che al variare dei mesi/stagioni l'effetto delle covariate sull'intensità degli incendi sia differente. Il problema è che anche raggruppando **Month** in **Season** (con 4 modalità) persiste una diversa esposizione delle variabili tra le diverse stagioni. E' chiaramente visibile il problema delle sovra/sotto esposizioni nei boxplot seguenti:

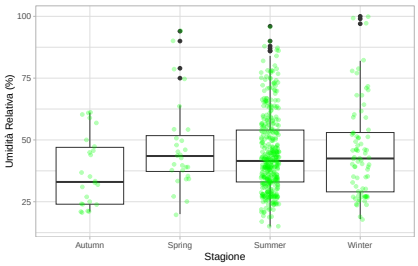


Figure 2: Umidità relativa per stagione

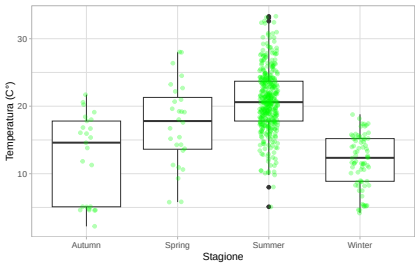


Figure 3: Temperatura per stagione

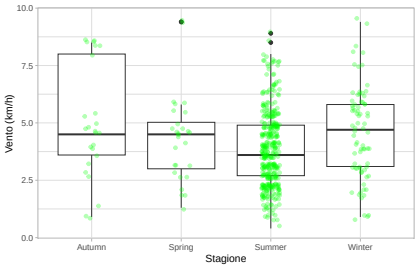
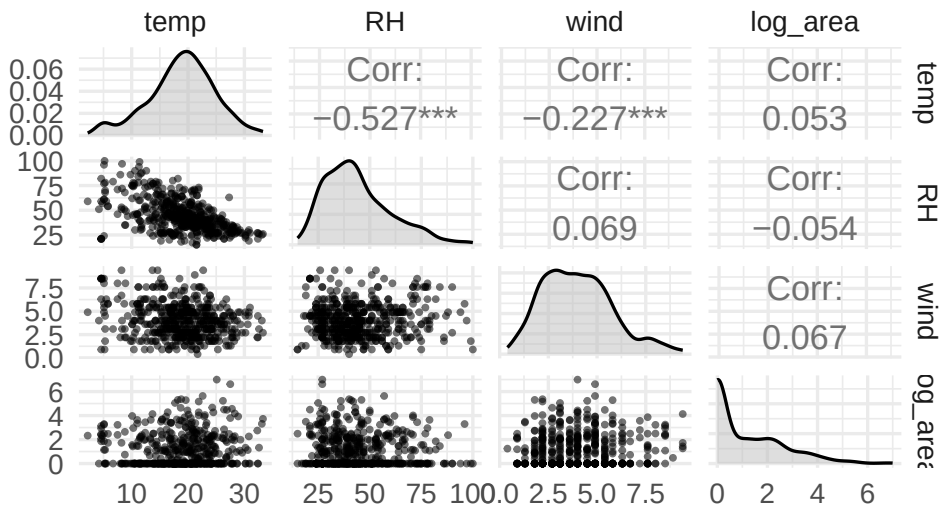


Figure 4: Velocità del vento per stagione

Volendo evitare di dover anestetizzare il modello per problemi di rappresentazione e fare retrospective sampling, abbiamo deciso di non usare **Day** o **Month** (o loro raggruppamenti) come etichette, ma di stimarle noi stessi con MoE. Chiaramente si è poi rivelata decisamente più complicata la fase di caratterizzazione dei gruppi (i quali non avranno più nulla a che fare con le stagioni o periodi della settimana). Le variabili esplicative scelte alla fine sono state: **Temp**,**Wind** e **RH**; in quanto facili da interpretare e adatte al modello usato. Abbiamo quindi analizzato le correlazioni e gli andamenti marginali delle variabili selezionate.



Marginalmente sembrano non esserci chiari andamenti lineari, è plausibile invece la presenza di andamenti lineari sovrapposti per via dei cluster. Inoltre si può notare come le correlazioni siano molto basse tra la risposta trasformata e le esplicative; Nonostante ciò procediamo con il modello di regressione di misture, aspettandoci che le future correlazioni nei sottogruppi stimati saranno maggiori. Sulla diagonale invece sono riportate le densità non-parametriche stimate con Kernel gaussiano e bw di default. Si notino le multimodalità sulle variabili esplicative ,le quali suggeriscono la presenza di cluster latenti.

2.3 Mixture of Experts model

In questa sezione cerchiamo di rispondere alla seguente domanda di ricerca: Esiste una variabile latente che differenzia l'impatto che le variabili **temp**, **wind** e **RH** hanno sull'intensità degli incendi? L'intensità (**log_Area**) è misurata come trasformata logaritmica della area bruciata. Come prima cosa abbiamo centrato e standardizzato i dati in modo da favorire la convergenza degli algoritmi iterativi per la stima del modello. Abbiamo successivamente scartato diversi gating network MEM e expert network MEM per via del loro elevato ICL. Infine abbiamo scelto un modello Full MEM, nel quale le covariate **Temp**, **Wind** e **RH** entrassero a spiegare sia le mixing proportions che le medie di gruppo. Di seguito le caratteristiche tecniche del modello utilizzato (le quantità in grassetto sono da intendersi come vettori):

- abbiamo optato per una regressione di misture di K normali univariate.
- i K GLM per modellare le medie di gruppo sono stati semplificati in LM mentre per stimare le mixing proportions è stata usata una regressione multinomiale logistica a K categorie
- La funzione di densità della mistura di K normali da stimare è la seguente: $f_{Y_i|\mathbf{x}_i}(y_i \mid \mathbf{x}_i; \boldsymbol{\eta}) = \sum_{j=1}^{K} p_j(\mathbf{x}_i)\, f_j\big(y_i; \mu_j(\mathbf{x}_i)\big)$ le f sono da intendersi come densità di normali univariate.
- j-esima componente della regressione multinomiale logistica: $p_j(\mathbf{x}_i) = \frac{\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}_j\right)}{1+\sum_{j=1}^{K-1}\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}_j\right)}$ dove $\mathbf{x}_i^\top \boldsymbol{\beta}_j = \beta_{j0} + \beta_{j1}\,\mathrm{Wind}_i + \beta_{j2}\,\mathrm{Temp}_i + \beta_{j3}\,\mathrm{RH}_i$
- j-esimo LM : $\mu_j(\mathbf{x}_i) = \gamma_{j0} + \gamma_{j1}\,\mathrm{Wind}_i + \gamma_{j2}\,\mathrm{Temp}_i + \gamma_{j3}\,\mathrm{RH}_i$

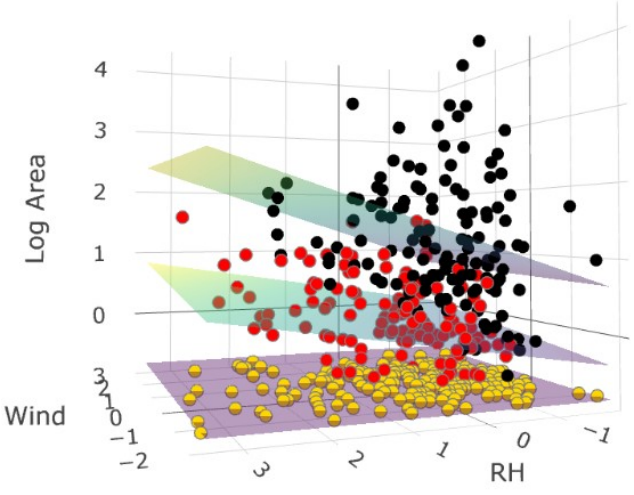
Per stimare il modello è stata usata la funzione stepFlexmix dal pacchetto Flexmix di R, la quale ci ha permesso di scegliere, tra i modelli migliori in termini di ICL, quello che ci sembrasse più interpretabile. Abbiamo quindi stimato 5 modelli per K da 2 a 4. Il numero di cluster ci sembrava ragionevole porlo da 2 a 4. Questa scelta è stata in parte influenzata dall'interpretazione ,poi scartata, che i cluster fossero determinati dalle stagioni, ma soprattutto dal numero restrittivo di osservazioni. Abbiamo infine scelto un modello mistura a 3 componenti sulla risposta con mixing proportions e medie di gruppo spiegate dalle 3 covariate già menzionate.

2.4 Risultati

Table 1: Coefficienti stimati dal modello Full MEM selezionato

	LM coeff			**Multi Logit Reg. Coeff**		
Variable	$\underline{\gamma}_1$	$\underline{\gamma}_2$	$\underline{\gamma}_3$	$\underline{\beta}_1$	$\underline{\beta}_2$	$\underline{\beta}_3$
coef.(Intercept)	0.1219	1.198	-7.945e-01	0	0.2080	0.8302
coef. **Temp**	-0.2790	-0.1262	-3.578e-17	0	0.9052	0.2199
coef. **RH**	-0.06276	-0.1406	-1.290e-17	0	0.1834	0.01865
coef. **wind**	0.04645	0.2426	2.130e-17	0	-0.4726	-0.3639
sigma	0.4194	0.9152	2.195e-16	–	–	–

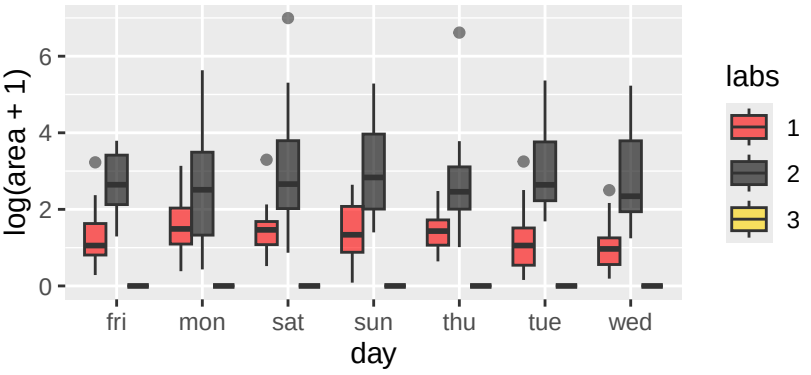
AIC: -15880.37 **BIC:** -15782.66 **ICL:** -15671.85 ***Avg.Unc.*** = 0.088



In alto si possono osservare tabulati i coefficienti stimati degli LM e della regressione multinomiale logistica. Si noti l'impatto diversificato tra i gruppi delle esplicative sulla risposta (sintetizzato nei coefficienti di regressione). A Sinistra si può invece vedere l'impatto diversificato per cluster che le variabili **RH** e **Wind** hanno sulla **log_area**. Si notino i piani di regressione in trasparenza. Abbiamo scelto arbitrariamente di rappresentare solo **RH** ,**Wind** ma ovviamente il modello permetterebbe di plottare tutte le combinazioni possibili di esplicative prese a coppie di due per spiegare la risposta. I colori delle nuvole di punti sono stati assegnati in base all'etichetta stimata dal modello Full MeM. La colorazione è stata scelta in modo evocativo a richiamare diversi livelli di intensità dell'incendio. Ricordiamo che le esplicative sono 3 e questo è un grafico parziale.

2.5 Interpretazione dei risultati

In questo primo grafico si può osservare ,tramite boxplot condizionati all'etichetta stimata, l'andamento dell'intensità degli incendi nei diversi giorni della settimana. Ciò ha senso dal momento che la maggior parte di incendi sono di causa antropica e il parco Naturale è una meta turistica soggetta a flussi di visitatori non costanti nella settimana. Si noti che i boxplot gialli sono sempre presenti ma schiacciati sullo 0. Lo stesso grafico si potrebbe fare con i mesi sulle ascisse. Si nota che non vi sono grandi differenze tra giorni della settimana, il che ha rafforzato l'idea che clusterizzare rispetto a **Day** sarebbe risultato fallimentare



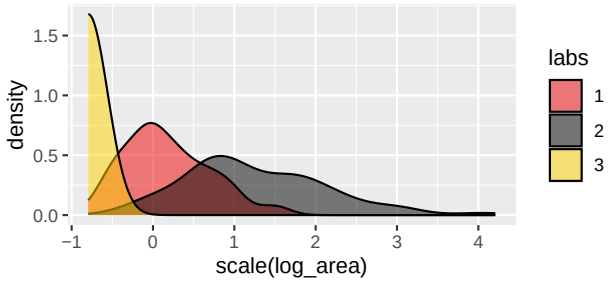


Figure 5: In questo secondo grafico abbiamo le stime kernel della risposta condizionate ai tre gruppi stimati. Si noti la diversa variabilità dei tre gruppi e il fatto che venga ricostruita la mistura sulla risposta

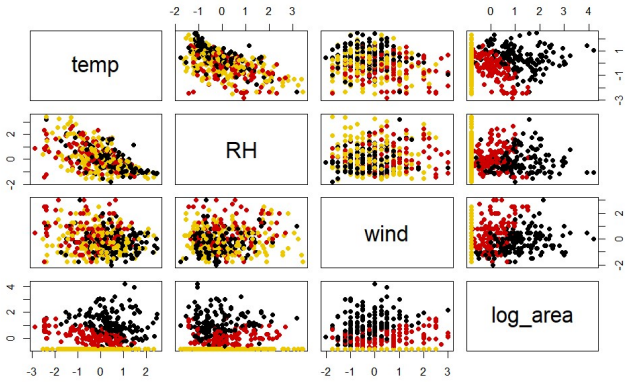
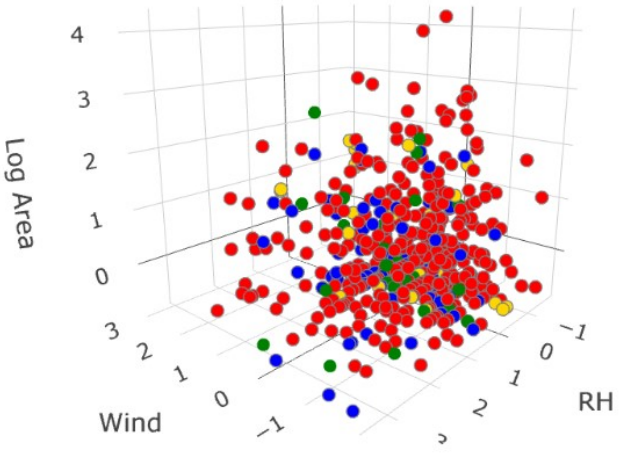


Figure 6: Questo terzo grafico presenta le proiezioni bidimensionali di covariate e risposta per ciascuna coppia di variabili. I colori sono assegnati in base alle etichette stimate.

Il grafico 6 si presta particolarmente bene a commenti di natura interpretativa. Si può notare infatti come la clusterizzazione che è stata fatta rispetto alla variabile risposta **log_area** porti a 3 gruppi sui quali l'impatto delle variabili metereologiche sull'area bruciata è differente. A livello marginale, come ci aspettavamo, per valori di **RH** bassi si verificano gli incendi più intensi, lo stesso accade quando **Temp** è elevata.

A posteriori siamo in grado di dire che la clusterizzazione stimata da MEM non si ripresenta se dividiamo le osservazioni in base a **Season** (raggruppamento di **Month** in 4 classi, nella figura a destra) o a raggruppamenti di **Day**. Non solo non si ripresenta ma non si creano cluster evidenti che facciano propendere per un effetto diversificato sulla risposta. Pertanto siamo in grado di conludere che l'effetto differenziato delle variabili metereologiche sugli incendi non dipende ne dal periodo settimanale ne da periodi annuali/stagioni. Risulta evidente nel grafico a destra (anche se sarebbero necessarie ulteriori analisi). Si noti che anche cambiando le coppie di esplicative non si presentano cluster significativi (o simili a quelli stimati con MoE).



3 Classification

In questa seconda sezione del Report mostriamo i risultati inerenti alla Classification. In modo particolare abbiamo deciso di considerare come Ground-truth le etichette stimate in modo unsupervised dal modello Full MEM. Abbiamo poi deciso di utilizzare un classifier di tipo EDDA e implementare la fase di learning e selezione del classificatore (sul training set) tramite la funzione mixModLearn della library Rmixmod di R. In questo contesto, abbiamo deciso di utilizzare come features per la classification le sole variabili esplicative. Il classifier è stato scelto reiterando 1000 volte la mixmodlearn e scegliendo il modello che minimizzasse il Mer. Di seguito è riportato l'andamento della stima CV 10-fold del Mer per modelli con Mixture Sampling.

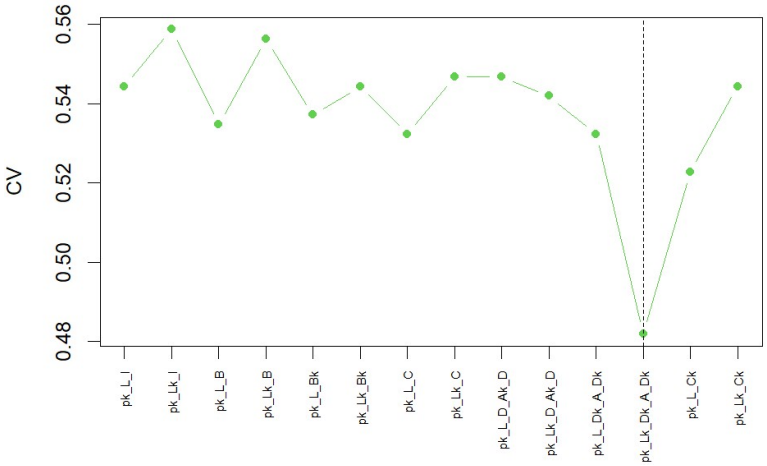


Figure 7: il MER minimo = 0.4820. Sulle ascisse i modelli in ordine di complessità crescente da sx a dx

Il classifier migliore in termini di CV risulta il modello pk_Lk_Dk_A_Dk ossia una mistura di 3 normali trivariate con vincolo di tipo VEV sulla struttura di variabilità (in notazione VSO). In altra parole abbiamo scelto un modello senza vincoli su volume dei cluster e orientamento, ma con shape dei cluster uguale. Il classifier vero e proprio pertanto sarà di tipo optimal bayes basato sulla mistura di normali appena descritta. Abbiamo deciso di non usare la variabile risposta come feature in quanto già utilizzata per fare clustering. Pertanto ci sembrava più sensato usare solo le esplicative per fare classification.

Di seguito sono riportate le misure di specificità e sensitività del classifier allenato e testato sul test set costituito da 100 osservazioni e la Confiusion matrix. Si noti che il classificatore produce un accuracy di 0.63.

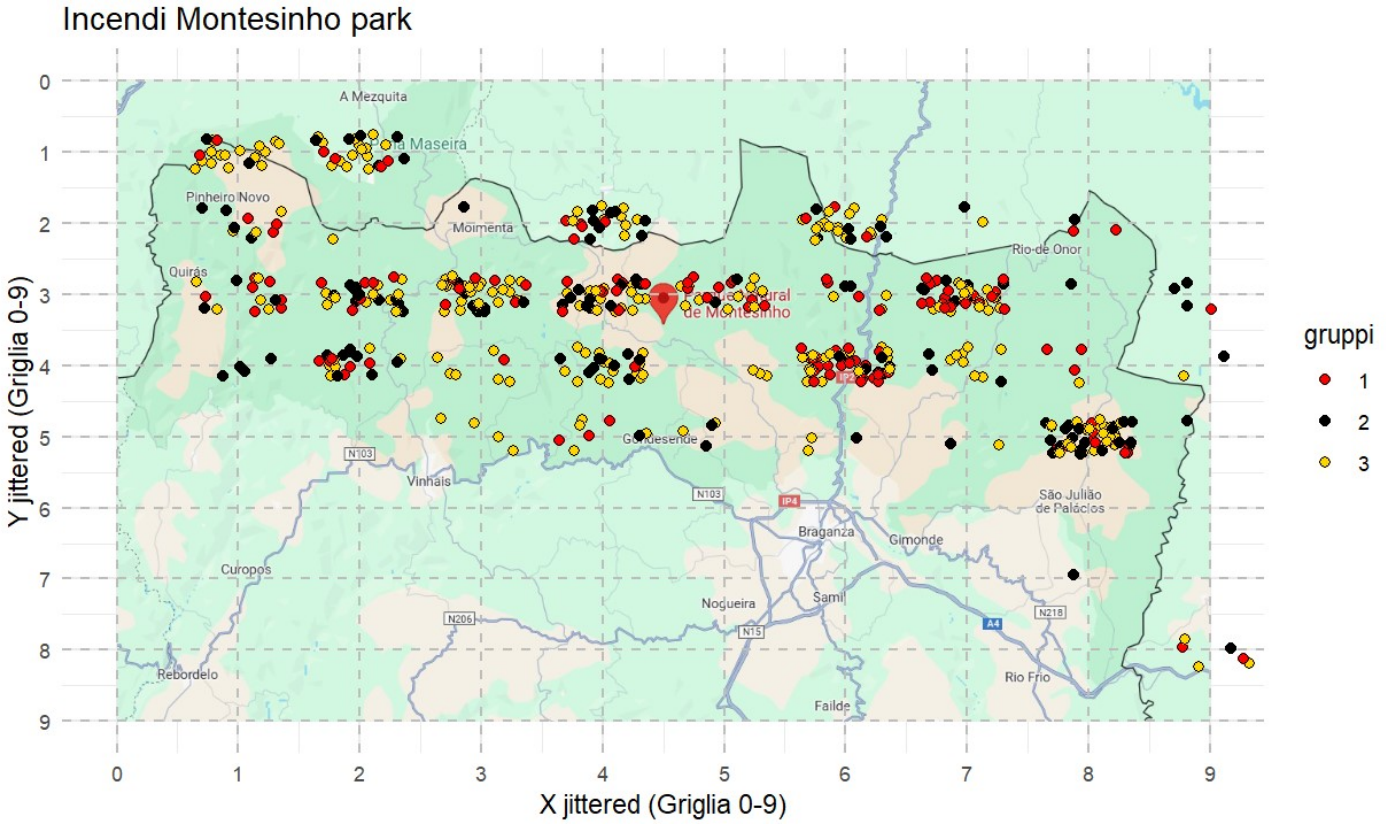
	Class 1	Class 2	Class 3
Sensitivity	0.4615	0.5161	0.7321
Specificity	0.9425	0.8261	0.5455

Confusion Matrix			
Ground-truth	1	2	3
Predicted 1	6	0	5
Predicted 2	2	16	10
Predicted 3	5	15	41

Table 2: Sensitività, specificità e Confusion matrix.

4 Ipotesi sulla caratterizzazione dei gruppi

La scelta di intraprendere un'analisi statistica di tipo unsupervised ci impone o quanto meno richiede uno sforzo ulteriore al fine di caratterizzare i gruppi trovati. Durante l'analisi abbiamo scartato le ipotesi apparentemente più plausibili di cluster differenziati in base a **Month** (o suoi raggruppamenti) e **Day**. Abbiamo quindi deciso di cambiare framework provando a sfruttare le scarse informazioni geografiche presenti nel dataset . Volevamo comprendere se la localizzazione approssimata dell'incendio potesse influenzare l'effetto delle covariate sull'area bruciata. In modo particolare la nostra intuizione è stata di provare a combinare le coordinate geografiche degli incendi (**X** e **Y**) con l'altituidine alla quale si sono verificati (deducibile dalla mappa). In questo contesto l'idea era di capire se l'effetto diversificato delle covariate sulla risposta fosse dovuto all'altitudine (e quindi alla concentrazione diversa di Ossigeno), o più in generale a una combinazione di fattori geografici. E' importante osservare che i dati a disposizione sono profondamente distorti. Le vere coordinate degli incendi sono state (per motivi vari) approssimate al punto più vicino della griglia 9x9 sovraimposta alla mappa. E' chiaro quindi che 2 livelli di approssimazione ,ossia, i punti rispetto alla griglia e la griglia rispetto alla mappa inducono una pesante distorsione sui dati. Consapevoli di ciò abbiamo comunque provato a vedere se la clusterizzazione stimata con Full MEM combinata con le coordinate ci potesse suggerire una caratterizzazione dei cluster.



4.1 Conclusioni sulle ipotesi

La nostra ipotesi a priori era che incendi allocati nel cluster giallo (bassa intensità) fossero prevalenti ad altitudini piu elevate (per la carenza di ossigeno) mentre quelli neri (più intensi) fossero più frequenti in pianura, e i rossi una via di mezzo. In alto è riportata la mappa del Montesinho Natural Park che abbiamo realizzato per rispondere alla nostra ipotesi. Si noti che abbiamo introdotto in modo arbitrario del rumore tramite jitter, alternativamente avremmo ottenuto una rete poco informativa di punti sui nodi della griglia. Nonostante il grafico non sia molto attendibile, ci sembra di poter escludere che l'altitudine o particolari fattori geografici influenzino i cluster stimati. Una seconda ipotesi era che la vicinanza ad un centro abitato di medie dimensioni si traducesse in una risposta più veloce a domare il fuoco. Si noti dalla mappa che il numero e la gravità degli incendi sembra aumentare quando ci si allontana dalla città di Braganza e Vinhais, quasi a suggerire la presenza di due cluster centrati nelle rispettive cittadine. Ciò non è però supportato dai dati visto che il modello migliore (e più interpretabile) ci suggerisce 3 classi. Possiamo quindi affermare che il problema di interpretazione dei gruppi rimane aperto dal momento che non siamo riusciti a caratterizzare le classi in base alla geografia degli incendi. Rimandiamo questo problema a futuri lavori.