

## Enunciat 2 – Models per comptatges

L'objectiu serà construir diversos models generalitzats pel recompte de casos de avvistaments d'ovnis en Estats Units en aquest mil·lenni. Les dades han estat recopilades pel National UFO Reporting Center d'Estats Units (<http://www.nuforc.org/>) S'han considerat només avvistaments en els Estats Units des de l'any 2000 fins al 2014. Hi ha un total de 52.813 esdeveniments reportats. S'han classificats segons l'Estat, el quinquenni, el dia de la setmana, el mes de l'any i la franja horària. L'objectiu és determinar les condicions que estan relacionades amb un més gran valor esperat d'avvistaments. També es volen caracteritzar els Estats en base a la incidència i comprovar l'evolució al llarg d'aquest període.

*Només els millors models obtinguts seguint les indicacions haurà d'interpretar-se i justificar la seva bondat en termes d'explicabilitat, predicció i diagnòstic. Tanmateix caldrà jutjar de manera crítica l'adequació del model a la realitat de les dades.*

### Indicacions:

- Caldrà fer l'anàlisi exploratori preliminar unidimensional i descriure les relacions bivariants/multivariants observades segregant clarament els factors en relació amb la resposta.
- Es poden modificar les variables explicatives i crear de noves, com ara reagrupant categories.
- La variable Estat actuarà de forma additiva en els models, sense interaccionar amb la resta de factors. D'aquesta manera es podran caracteritzar fàcilment els Estats en base al coeficient corresponent.
- Es vol estudiar el tipus de dependència entre les variables categòriques, considerant interaccions fins a ordre 2.
- En la construcció del model cal tantejar només una funció d'enllaç (models loglineals).
- Tantegeu les diferents propostes de model probabilista per les respostes: **normal**, **lognormal**, **poisson**, **poisson sobredispersada (NB1)** i **binomial negativa (NB2)**.
- Podeu construir els models fent servir alguna tècnica per seleccionar les variables (eliminació de predictors no significatius, mecanismes stepwise a partir del model més complet)
- Discutir en el model resultant els efectes de les variables en la resposta. Fer una comparació entre els tres models en relació als predictors que han sortit significatius.
- El model resultant ha de ser interpretat en terme de les relacions dels predictors seleccionats i el seu efecte sobre la variable resposta (en l'escala del predictor resposta).
- Valoreu els indicadors de bondat de l'ajust.
- Comparar la distribució observada de la resposta amb la distribució predita pel model.
- Diagnòstic d'outliers dels residus, observacions influents, observacions d'anclatge elevat, valoració dels diagrames de residus, etc.

**Per lliurar ( feu-lo a la tasca corresponent del Campus Virtual o en cas de problemes per email a [josep.a.sanchez@upc.edu](mailto:josep.a.sanchez@upc.edu))**

- Informe en pdf on s'expliqui el procés de construcció, validació i interpretació del model final (extensió màxima: 12 pàgines). Pot definir-se en un annex (sense límit d'extensió) altres gràfics i sortides de R que es considerin interessants.
- Codi en R amb les instruccions que s'han fet servir

**Data límit de lliurament:** 9 de Gener de 2017 a les 23:59.

## Descripció de les dades

Nombre de avistament d'objectes volants no identificats en els Estats Units del 2000 al 2014

### Format

La matriu de dades disponible conté in formació de 50.813 avistaments agupats en 20.984 observacions i 6 variables:

<b>State</b>	Estat dels EEUU (50 categories)
<b>Period</b>	Quinqueni. (1:2000-2004, 2:2005-2009,3:2010-2014)
<b>Month</b>	Mes de l'any (1-12)
<b>Weekday</b>	Dia de la setmana (1:dilluns,...7:diumenge)
<b>Hora</b>	Franja horària (H00-05, H06-11, H12-17, H18-23)
<b>Sights</b>	Número d'avistaments