INRAE
Biostatistique
BIOSP
& Processus Spatiaux

AGENCE NATIONALE DE LA RECHERCHE
ANR

INRAE

Identifying factors impacting zero-inflated proportion data

Mélina Ribaud

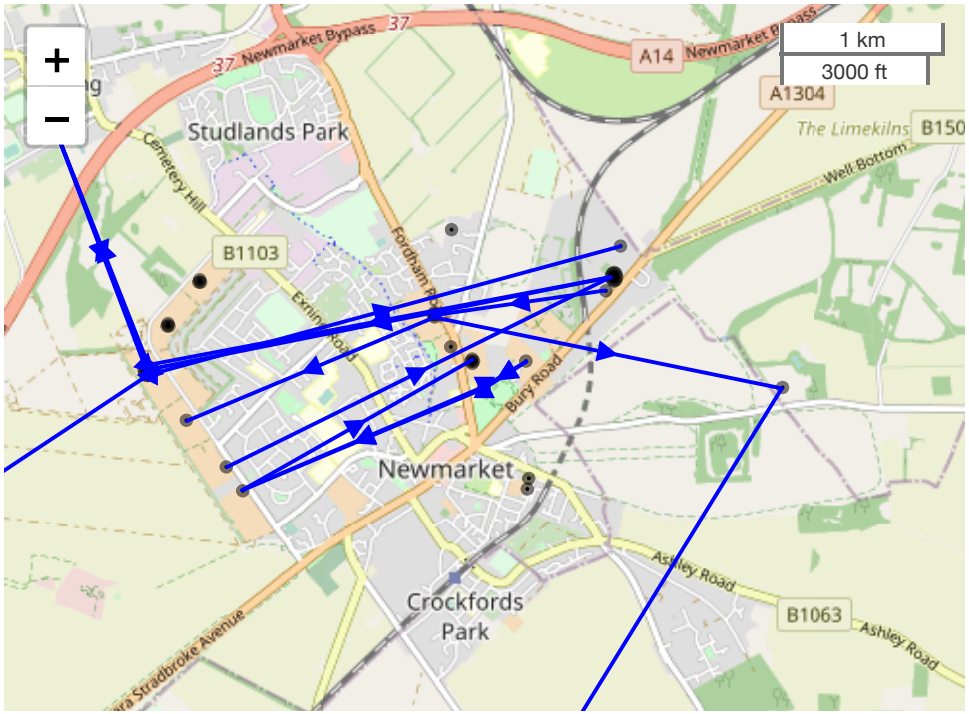
14/12/2020

Permutation Test, Rank Statistics, Performance Indicator

Equine Influenza, Covid-19, Canine rabies

Introduction

Data: Equine Influenza



Hughes et al. (2012)

Data preparation

Equine Influenza:

- $m = 47$ contributors (sources)
- $n = 47$ targets (receivers)

$$Z_j^i = \begin{cases} p & \text{if } j \to i \\ 0 & \text{if else} \end{cases}$$

Target blocs :

$$\sum_{j=1}^{n_c} z_j^i = 1$$

Response

Let Z_j^i be a random variable associated with the target $i \in \{1, \dots, n_t\}$ and the contributor $j \in \{1, \dots, n_c\}$.

We assume that :

- Z_j^i is continuous,
- $Z_j^i \in [0, 1]$,
- the distribution of Z_j^i is zero-inflated,
- for a fixed target node, the sum over all contributors cannot exceed 1, i.e.:

$$\sum_{j=1}^{n_c} Z_j^i \leq 1.$$

Factors

$\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_d) \in \mathbb{R}^{n_t n_c \times d}$

Equine Influenza: 1 *continuous factors (distance between hosts) and* 3 *discrete factors (Yard, Sex, Age and SexGender)*

In practice, factors often provide information about the target node i and the contributing node j separately, but not about the pair (i, j). In this case, $x_k^{(i,j)}$ can be defined from any application g:

$$\begin{array}{rcl} g : E \times E & \to & \mathbb{R} \\ (\mathbf{e}^i, \mathbf{e}^j) & \mapsto & g(\mathbf{e}^i, \mathbf{e}^j) = x_k^{(i,j)} \end{array}$$

Why a new method?

Table 1 Model comparison to match our constraints.

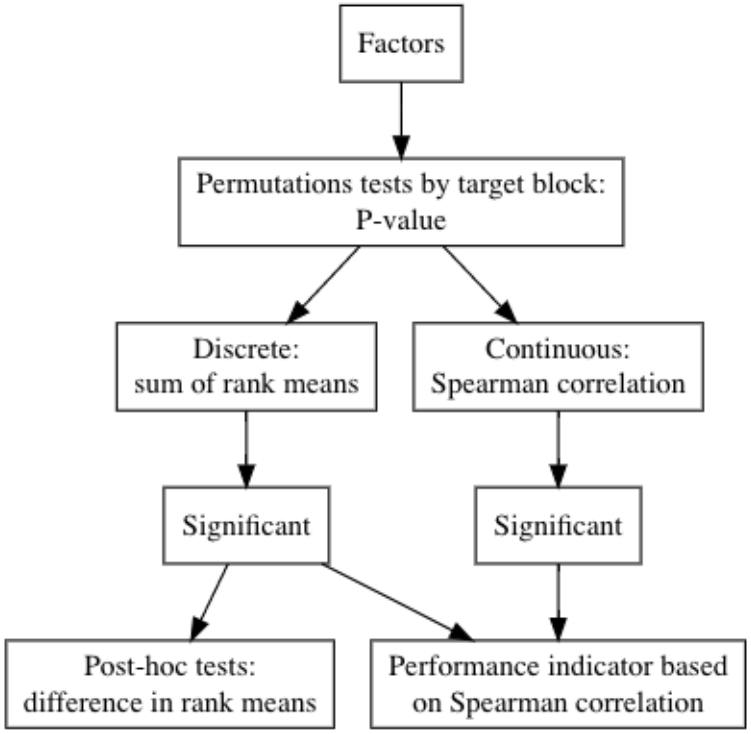
Methods	Response			Factor		Dependency
	No hypothesis on the law	$[0, 1]$	Zero inflated	Tests Discrete	Tests Continuous	
Linear regression (Hastie et al., 2009)				✓[a]	✓	
Beta regression (Stasinopoulos et al., 2007)		✓	✓	✓	✓	
Dirichlet regression (Tsagris and Stewart, 2018)		✓	✓	✓	✓	✓
Decision tree (Breiman et al., 1984)	✓	✓				✓

[a] ANOVA and ANCOVA

Hastie et al. (2009) Stasinopoulos. (2007) Tsagris and Stewart (2018) Breiman et al. (2009)

Method

Procedure



- *P-value (block CMC):*

$$\hat{\lambda}_k(\mathbf{z}) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\{T_k^b \geq T_k\}}$$

- *Statistic (discrete):*

$$T = (n_t n_c - 1)\frac{\sum_{q=1}^{Q} n_q \left(\bar{R}_{\mathbf{z}\cdot q} - \bar{R}_{\mathbf{z}}\right)^2}{\sum_{i=1}^{n_t}\sum_{j=1}^{n_c}(R_{z_j^i} - \bar{R}_{\mathbf{z}})^2}$$

- *Statistic (Continuous): Spearman correlation*

$$T = r_s^2(\mathbf{x}, \mathbf{z}) = \rho^2(R_{\mathbf{x}}, R_{\mathbf{z}})$$

Performance indicator

$$I_\beta(\mathbb{X}, \mathbf{z}) = r_s^2(M_\mathbb{X}\beta, \mathbf{z})(1 + \Delta_{M_\mathbb{X}\beta,\mathbf{z}})$$

Two optimizations

1.Maximization of $\left(\frac{1}{1+\Delta_{M_\mathbb{X}\beta,\mathbf{z}}} \right)$ *(upper bound) via genetic algorithm*

2.Maximization of $I_\beta(\mathbb{X}, \mathbf{z})$ *via genetic algorithm* (β)

Upper bound: maximum Spearman correlation that can be obtained by taking into account the structure of the response (multitudes of zeros and ties)

Results: Equine Influenza

Factor	T^*	pv	r_s
Yard	0.05	0	
Sex	0.007	0.031	
Age	0.001	0.8	
distanceYard	0.05	0	−0.22
SexGender	0.042	0	

Factor	Factor level	T^*	pv	pv*
Yard	0 - 1	-444	0	
Sex	0 - 1	-24.77	0.04	
SexGender	F->F - F->M	65	0	0.01
	F->F - M->F	111	0	0
	F->F - M->M	80	0	0
	F->M - M->F	46.2	0.01	0.02
	F->M - M->M	15	0.33	0.33
	M->F - M->M	-31.1	0.04	0.08

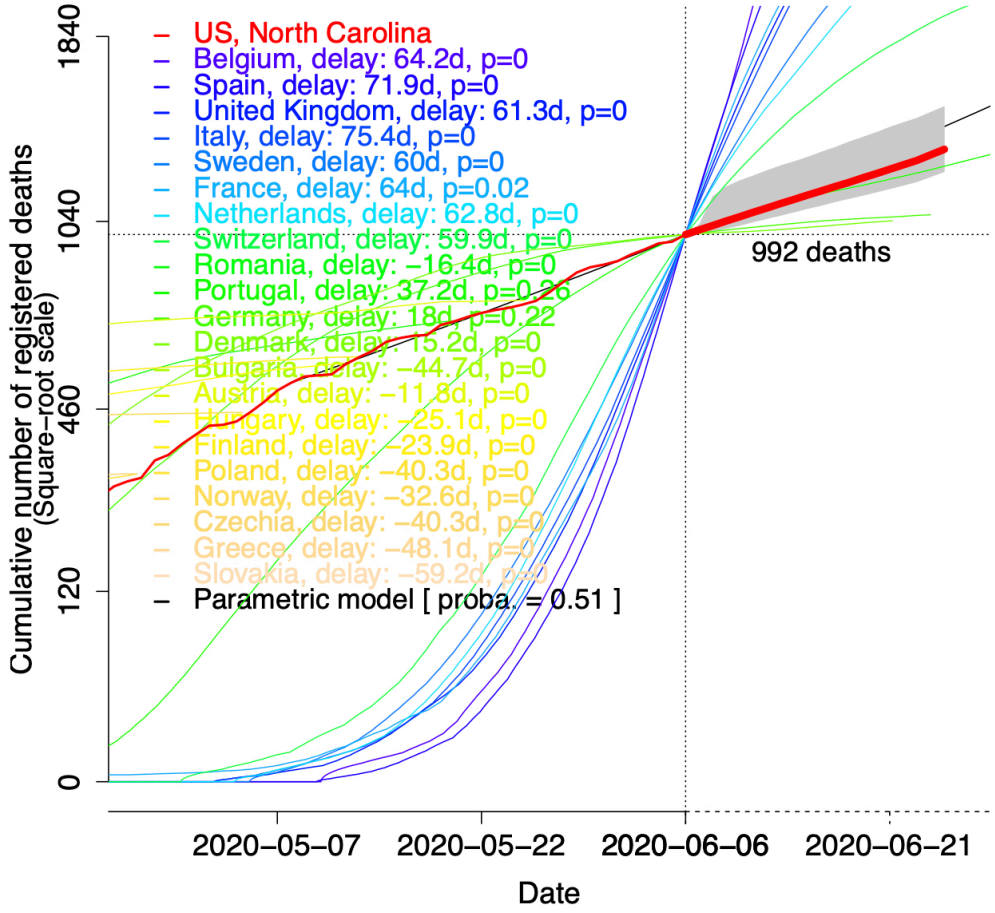
Female ==> “super” contaminators

$$I_{\hat{\beta}}(\mathbb{X}, Z) = 0.38$$

These factors explain only a little part of the entire correlation

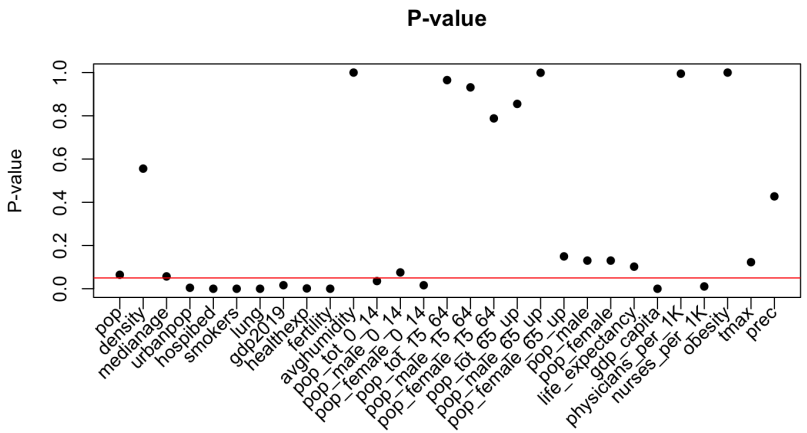
Covid-19 application

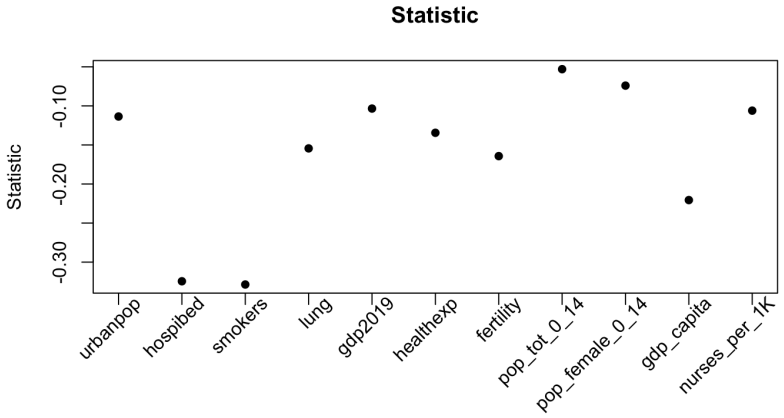
Data: 23 factors, 23 targets and 21 contributors



Soubeyrand et al. (2020) Shiny App

P-values, statistic, performance indicator and beta





Estimated Beta Values

	Beta
hospibed	*-2.70*
smokers	*-9.45*
lung	*-4.30*
healthexp	*-2.51*
gdp_capita	*-1.70*
fertility	*1.37*
urbanpop	*0.32*
nurses_per_1K	*-0.47*
gdp2019	*-0.93*
pop_female_0_14	*-2.19*
pop_tot_0_14	*6.10*

$$I_{\hat{\beta}}(\mathbb{X}, Z) = 0.78$$

Canine rabies

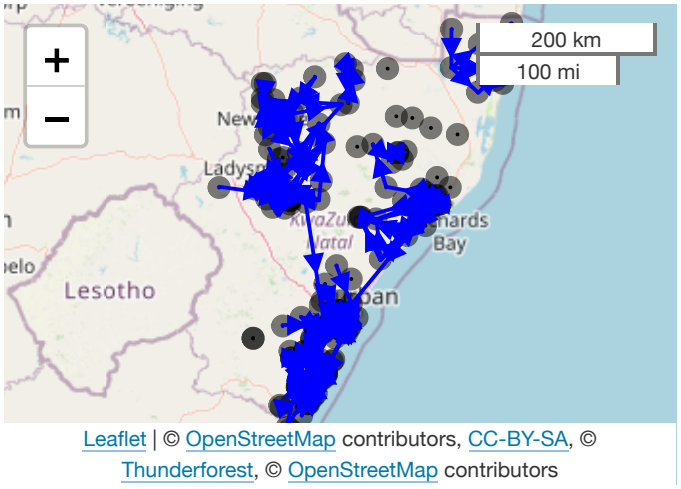
Data: 176 hosts and 24 factors

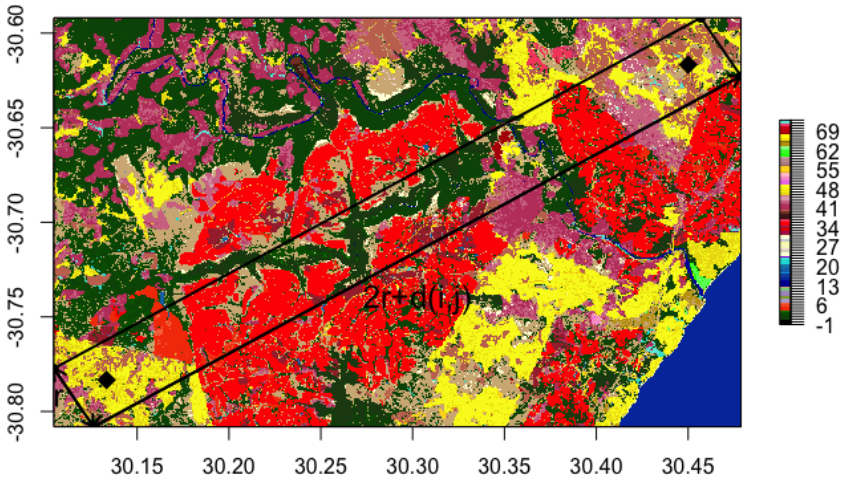
24 factors	
Natural Wooded Land	Temporary Crops
Planted Forest	Fallow Lands & Old Fields
Shrubs	Residential
Karoo & Fynbos Shubland	Village
Natural Grassland	Smallholdings
Natural Waterbodies	Urban Vegetation
Artificial Waterbodies	Commercial
Herbaceous Wetlands	Industrial
Woody Wetlands	Transport
Consolidated	Surface Infrastructure
Unconsolidated	Extraction Sites
Permanent Crops	Waste & Resource Dumps

More information pdf

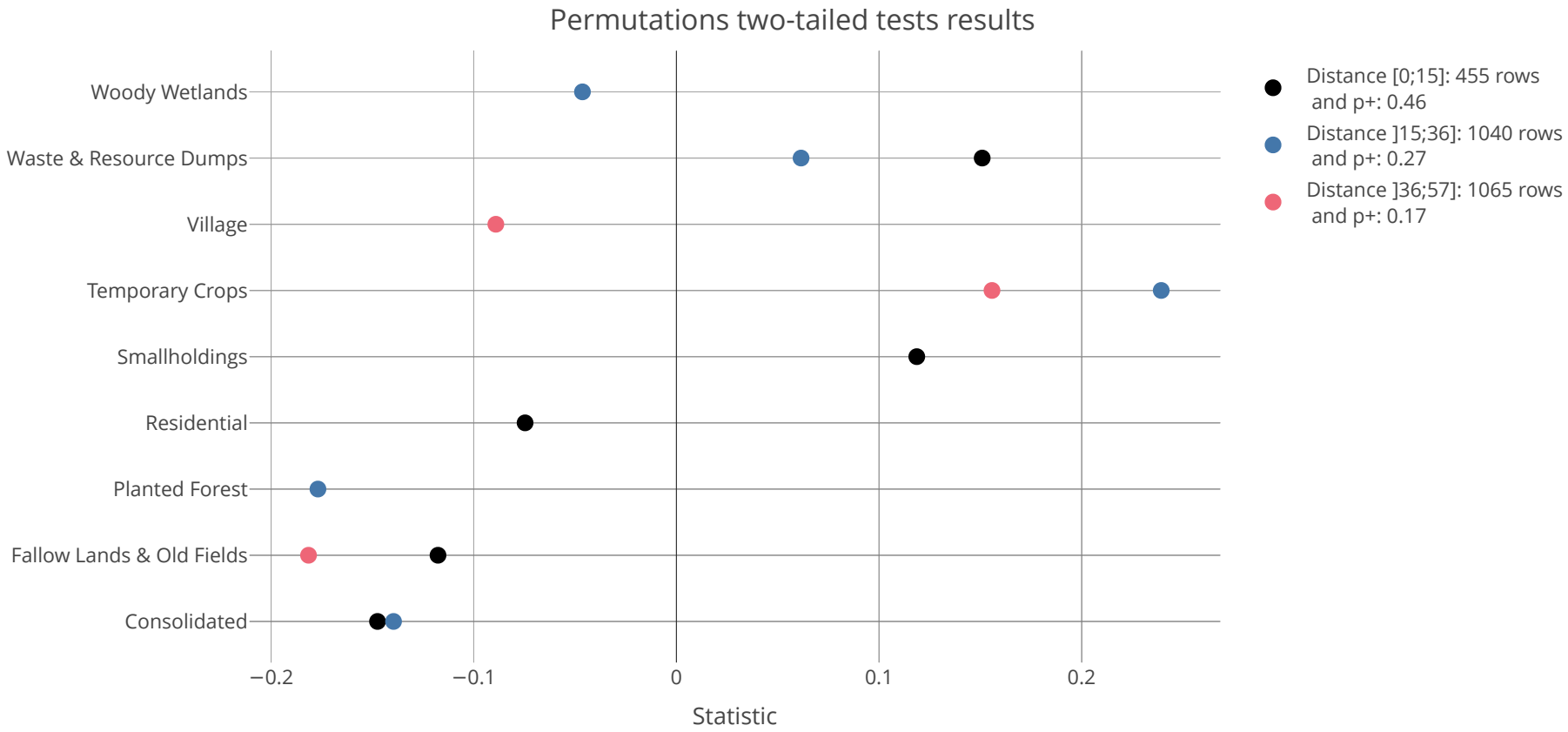
South African National Land-Cover (SANLC) 2018

Mollentze et al. (2014)





Results


Permutations two-tailed tests results
Woody Wetlands
Waste & Resource Dumps
Village
Temporary Crops
Smallholdings
Residential
Planted Forest
Fallow Lands & Old Fields
Consolidated
−0.2
−0.1
0
0.1
0.2
Statistic
Distance [0;15]: 455 rows and p+: 0.46
Distance]15;36]: 1040 rows and p+: 0.27
Distance]36;57]: 1065 rows and p+: 0.17

Conclusions

Perspectives

- Improve optimization of performance indicator
- Canine rabies

Material

- Pre-print HAL https://hal.archives-ouvertes.fr/hal-02936779
- Package ZIprop (Gitlab BioSP) https://gitlab.paca.inrae.fr/meribaud/ziprop