Identifying potential significant factors impacting zero-inflated proportions data

Melina Ribaud, Edith Gabriel, Joseph Hughes, Samuel Soubeyrand

IDENTIFYING POTENTIAL SIGNIFICANT FACTORS IMPACTING ZERO-INFLATED PROPORTIONS DATA

Melina Ribaud
INRAE, BioSP,
84914 Avignon, France
melina.ribaud@inrae.fr

Edith Gabriel
INRAE, BioSP,
84914 Avignon, France

Joseph Hughes
MRC-University of Glasgow,
Centre for Virus Research,
Glasgow, Scotland, United Kingdom

Samuel Soubeyrand
INRAE, BioSP,
84914 Avignon, France

September 29, 2020

ABSTRACT

Managing epidemics requires to investigate potential impact of risk and protective factors on epidemiological links. Here we focus on links defined by inferred probabilities (transmission links in Equine Influenza, similarity measures of COVID-19 dynamics between different countries). The specific nature of these epidemiological data (zero-inflated, correlated, continuous and bounded) does not allow to use classical supervised methods like linear regression or decision tree to identify impacting factors on the response variable. In this article we propose a by block-permutation-based methodology (i) to identify factors (discrete or continuous) that are potentially significant, (ii) to define a performance indicator to quantify the percentage of correlation explained by the significant factors subset. The methodology is illustrated on simulated data and on the above-mentioned epidemics.

***K*eywords** Permutation Tests; Spearman's correlation; Performance Indicator; Covid-19; Equine Influenza

1 Introduction

Effective strategies for the management of infectious diseases are of importance to prevent health crisis, as recently confirmed with the COVID-19 pandemic. Assessing the influence of social, biological and environmental factors in the spread of epidemics is the main purpose for levers identification to prevent and control any epidemics. Epidemiological links, e.g. probabilities of transmission, (i) have an intrinsic correlation structure and (ii) are usually estimated, making their relationship with different factors challenging. Here, we focus on epidemiological links defined by proportions and we aim to provide a statistical methodology to reduce the bias of estimation when explaining epidemiological links (hereafter, the response variable) by various potentially impacting[1] factors.

Many statistical methods can be used to identify the correlation between factors and response. Parametric prediction models can identify the set of factors impacting the response through statistical tests. When the response is normally distributed, or when data are transformed to make it fit a Gaussian distribution (Weisberg, 2005), the linear regression model (Hastie et al., 2009) predicts response values and identifies influencing factors. When the response variable follows another usual distribution (Binomial, Poisson ...), the generalized linear models (GLM) described in Nelder and Wedderburn (1972) can similarly be considered. When the distribution of the response variable cannot be found, non-parametric predictive models (Hastie et al., 2009) may be a solution. However, non-parametric models do not provide direct testing procedure to identify impacting factors.

[1]Note that if a factor impacts the response, it is then correlated with the response.

In the case of zero-inflated data, Estabrooks et al. (2004) introduce the so-called resampling methods for balancing classes. These methods are mainly used for categorical responses. For a continuous response, the model is often defined as a mixture of two processes: the first process generating zeros, the second process being governed by usual distributions; see for instance the definition of the zero-inflated Poisson, zero-inflated Beta or even zero-inflated Binomial distribution in Stasinopoulos et al. (2007). For such zero-inflated models, the influencing factors can also be identified via statistical tests.

The above-mentioned parametric models are defined for independent and identically distributed (iid) realizations. In many cases, proportions data are not independent as they sum to a fixed value (often one) and knowing their sum we can determine one proportion from the sum of the remainder.

In statistics, these data are referred to compositional data. Aitchison (1982) describe the mathematical framework of compositional data. Douma and Weedon (2019) propose a classification of the compositional data according to the nature of the response (proportions arising from counts vs from continuous measurements). Regarding the case of a zero-inflated continuous response, the Beta regression is a solution when the proportions work in pairs. The percentage of male and female for a given species is an example. When the observed categories are greater than two, the Dirichlet's regression is required. Tang and Chen (2019) propose the adaptation of the zero-inflated Dirichlet regression (ZIDR) model for microbiome compositional data.

Parametric methods provide statistical tests to quantify the significance of a factor which depend on the type of factors (discrete vs continuous). The statistical test provided by the linear model can treat all types of factors. ANOVA concerns discrete factors with more than 2 levels. The GLM (including zero-inflated data) and the ZIDR can treat continuous and discrete factors with only 2 levels.

In this article, we investigate the relationship between zero-inflated, non-Gaussian, correlated proportion data and several factors of any type. In epidemics caused by infectious diseases, several factors describing the environment, the habitat or the individuals influence the spread in hosts populations. Each potential source-receiver pair is described by different factors characterizing each of the two individuals or their interaction. The objective is to quantify the correlation between these factors and the probability that the related source-receiver pair is a real transmission. New statistical methods arise to infer transmission pathways from high throughput sequencing data; see e.g. Alamil et al. (2019) who developed a statistical learning approach for human, animal and plant diseases and Hughes et al. (2012) who inferred the transmission links in equine influenza. Epidemiological links can also be formalized by similarity measures between epidemic dynamics. Soubeyrand et al. (2020b) inferred the probability at a given time that a focal country follows the mortality trajectory of a benchmark country and applied the methodology to COVID-19 epidemics in Soubeyrand et al. (2020a).

The objective is now to investigate the impact of environmental, economical, climatic, ..., factors on epidemiological links. In both cases, data contains receivers and sources and the response variable represents inferred probabilities. The response is continuous and bounded by 0 and 1 with many zeros. The sum of probabilities by receiver block is less than or equal to one. The number of categories is equal to the number of sources. Usually, the number of potential sources are greater than two. Motivated by equine influenza and COVID-19 epidemics, our methodology aims (i) to identify factors (discrete or continuous) that are potentially significant and which can describe the pair of source-receiver or respectively the source and the receiver, (ii) to define a performance indicator to quantify the proportion of correlation explained by the significant factors subset.

The structure of the data and the objectives generate constraints on statistical modeling. The response takes values between 0 and 1 (inclusive) and is zero-inflated. Consequently, classical transformations to make it fit a Gaussian distribution can not be applied. In addition, the realizations are dependent due to the constraint on the sum of probabilities. Hence, linear regression which assumes a normal distribution is not appropriate. The beta regression described in Stasinopoulos et al. (2007) can be used to solve the non-normality constraint but the dependency structure is not taken into account. The ZIDR described in Douma and Weedon (2019) could be a solution to this dependency constraint. However, this method is not implemented for zero-inflated response and assumes that factors are fixed within categories. In our cases, a categories represents a block of receivers and each pair source-receiver has an own factor value. Consequently, factor varies within categories. For these four methods, a performance indicator can be defined thanks to a quality criterion like R2, RMSE or others. Table 1 summarizes the abilities of each method to match with our data constraints.

In this article, we propose a model-free approach, based on permutation tests, to identify influencing factors. Permutations tests (Pesarin and Salmaso (2010)) are widely used in biology; for instance Segal et al. (2018) propose a fast approximation of small p-values in permutation tests and Shih and Fay (1999) introduce a class of permutations tests for stratified survival data. Here, the permutations are constrained by the dependence structure. The test statistic depends

Table 1: Model comparison to match our constraints.

Methods	Response		Factor			Dependency
	$[0, 1]$	Zero-inflated	Varies within categories	Tests		
				Discrete	Continuous	
Linear regression			✓	✓[a]	✓	
Beta regression	✓	✓	✓	✓[b]	✓	
Dirichlet regression	✓	✓		✓[b]	✓	✓
Regression tree	✓		✓			✓

[a] ANOVA and ANCOVA
[b] Limited to 2 levels

on factor's type. If the factor is discrete, then the statistic is defined as the mean of the response by factors level. If the factor is continuous, then the statistic is the Spearman's correlation (see e.g. Hauke and Kossowski (2011)). The sign of the statistic value gives the factor effect. Factors with a significant correlation are then used to define a performance indicator. This indicator is an original proposition based on the Spearman's correlation. It quantifies the percentage of correlation explained by the selected set of factors. Figure 1 presents each step of the procedure.

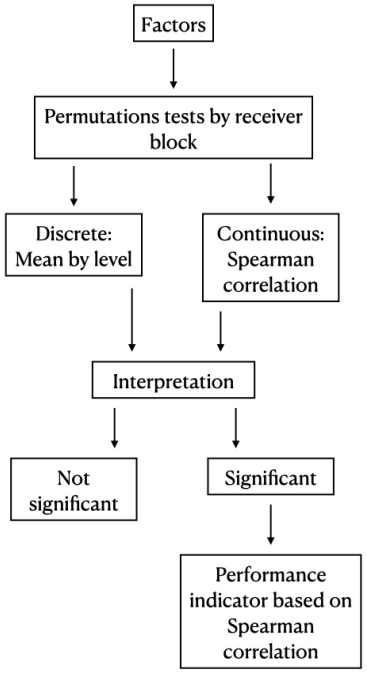

Figure 1: Procedure to identify correlation sign, significant correlation and to compute a performance indicator.

The remainder of this article is as follows. Framework and notations are set in Section 2. Section 3 introduces a procedure based on permutation tests to identify the factors correlated to the response and the performance indicator. The method is then illustrated on simulations (Section 4) and equine influenza and COVID-19 epidemics (Section 5).

2 Framework and notations

This section presents the framework of the method, linking epidemiological terminology and statistics and setting the structure of the response variable and factors.

Most of infectious diseases are transmitted by viruses intra or inter hosts populations. In epidemiology, a source refers to the person, animal, plant, ..., from which an infectious agent passes to the host. The receiver contracts the disease from a source. Hereafter, we consider that we have n_r receivers and n_s sources.

2.1 Response variable

Let Z_j^i be a random variable associated with the receiver $i \in \{1, \dots, n_r\}$ and the source $j \in \{1, \dots, n_s\}$. This variable defines the epidemiological link, i.e. the response variable. We assume that

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

$$\sum_{j=1}^{n_s} Z_j^i \leq 1 \tag{1}$$

For our epidemics, the number of strictly positive response values represents 10 to 30 percent of the dataset.

2.2 Set of factors

Usually, diseases are more likely to occur in some individuals of a population than others because of factors that may not be distributed randomly in the population. Hence, as noted earlier, most of epidemiological studies aim at identifying the risk and protective factors that place some individuals at greater and lesser risk than others. Here, factors describe a source-receiver pair or its interaction.

We denote by $\mathbb{X} = (\mathbf{x}_1, \dots, \mathbf{x}_d) \in \mathcal{M}_{n_r n_s \times d}$ the set of d factors that describes all pairs of source-receiver. Thus, any pair (i, j) is described by $\left(x_1^{(i,j)}, \dots, x_d^{(i,j)}\right)$. Often in practice, factors provide information about the receiver i and the source j separately, but not about the pair (i, j). In this case, we define $x_k^{(i,j)}$ by the distance (difference) between the receiver and the source:

$$x_k^{(i,j)} = |x_k^i - x_k^j|, \ \forall k \in \{1, \dots, d\}. \tag{2}$$

3 Identification and quantification of impacting factors

We propose a general methodology to identify factors that are correlated to the response. First, we introduce a permutation-based approach to identify the most significant factors. Then, we build an optimal performance indicator that quantifies the proportion of correlation explained by the selected factors.

3.1 A permutation-based approach to identify influencing factors

The specific characteristics of our response variable make the use of classical correlation tests (see e.g. Hollander et al. (2013) for Spearman test) impossible. Indeed, the response has numerous ties (zeros). Kendall (1945) proposes a solution to treat ties in ranking problems. When the ties are numerous, some hypothesis on the moments have to be satisfied. This verification could be laborious. Furthermore, the response is dependent by block of receivers (cf Equation(1) and classical correlation tests do not take into account such dependence structure. Consequently, by-block-permutation tests are a good alternative to take into account the constraints.

Let $\mathbf{x}_k \in \mathbb{R}^{n_s n_r}$, $k = 1, \dots, d$, be the observations of the factor to be tested and $\mathbf{z} \in \mathbb{R}^{n_s n_r}$, be the observations of the response. We use the Conditional Monte Carlo (CMC) algorithm described in Pesarin and Salmaso (2010) to test H_0: "the response is not correlated with the factor" versus H_1: "the response is correlated with the factor". We denote T the

statistics of test, which depends on the type of the factor and is defined below, and $\lambda_T(\mathbf{z}) = \mathbb{P}(T \geq T^*)$ the p-value.

A conditional Monte Carlo algorithm for block-permutation test:

1. Compute the statistic T^* on the original data set $(\mathbf{x}_k, \mathbf{z})$.
2. Randomly permute the response by block of receivers and define a new response vector denoted $\mathbf{z}^{\pi^1}$. Compute the statistic T^{π^1} on the permuted dataset $(\mathbf{x}_k, \mathbf{z}^{\pi^1})$.
3. Do B independent repetitions of step 2.
4. Estimate the p-value by $\hat{\lambda}_T(\mathbf{z}) = \frac{1}{B}\sum_{l=1}^{B} \mathbf{1}_{\{T^{\pi^l} \geq T^*\}}$.

Note that block permutations are required to get the most powerful the test (see Appendix A.1.).

For a discrete factor $\mathbf{x}$[2] with Q levels, we have $\frac{Q(Q-1)}{2}$ statistics defined by:

$$T = \frac{1}{n_r n_s \hat{\sigma}^2_{\mathbf{z}}} \sum_{q=1}^{Q} n_q \left(\bar{\mathbf{z}}_{.q} - \bar{\mathbf{z}}\right)^2, \tag{3}$$

where q are the levels of $\mathbf{x}$, $\bar{\mathbf{z}} = \frac{1}{n_r n_s}\sum_{i=1}^{n_r}\sum_{j=1}^{n_s} z^i_j$, $\hat{\sigma}^2_{\mathbf{z}} = \frac{1}{n_r n_s}\sum_{i=1}^{n_r}\sum_{j=1}^{n_s}(z^i_j - \bar{\mathbf{z}})^2$, $n_q = \sum_{i=1}^{n_r}\sum_{j=1}^{n_s}\mathbf{1}_{\{x^{(i,j)}=q\}}$ and $\bar{z}_{.q} = \frac{1}{n_q}\sum_{i=1}^{n_q}\sum_{i=1}^{n_s} z_{iq}$ ($z_{iq} = z^i_j$ and exists if $x^{(i,j)} = q$).

For a continuous factor $\mathbf{x}$, the statistics T is defined from the non-parametric Spearman's correlation, $r_s(\mathbf{x}, \mathbf{z})$, because its relation with the response is often not linear and the response variable is not normally distributed. The Spearman's correlation is defined as the Pearson's correlation between the rank variables (Spearman (1904)). Hence, we define

$$T = r_s^2(\mathbf{x}, \mathbf{z}) = \rho^2(R_{\mathbf{x}}, R_{\mathbf{z}}), \tag{4}$$

where $R_{\mathbf{x}}$ is the random variable that represents the rank of $\mathbf{x}$ such that $R_{x^{(i,j)}}$ is the rank of $x^{(i,j)}$ in $\left(x^{(1,1)}, \ldots, x^{(n_s,n_r)}\right)$; $R_{\mathbf{z}}$ is the random variable that represents the rank of $\mathbf{z}$ and ρ is the Pearson correlation.

3.2 A performance indicator to quantify the part of the correlation explained

In the previous section, only factors that are individually correlated to the response can be detected. However, the multivariate aspect is not taken into account. Here we deal with this multivariate aspect by developing a performance indicator that takes into account all discrete and continuous factors previously selected. Our indicator is defined as the ratio between the Spearman correlation and the optimal Spearman correlation, i.e. the maximum correlation given the large number of zeros: $I_{\boldsymbol{\beta}}(\mathbb{X}, \mathbf{z}) = r^2(\mathbb{X}\boldsymbol{\beta}, \mathbf{z})/r^2_{opt}(\mathbb{X}\boldsymbol{\beta}, \mathbf{z})$.

Performance indicator:

$$I_{\boldsymbol{\beta}}(\mathbb{X}, \mathbf{z}) = r_s^2(M_{\mathbb{X}}\boldsymbol{\beta}, \mathbf{z})(1 + \Delta_{M_{\mathbb{X}}\boldsymbol{\beta}, \mathbf{z}}), \tag{5}$$

where the elements of $M_{\mathbb{X}} \in \mathcal{M}_{(n_s n_r)} \times p$ are defined by:

$$M_{\mathbb{X}}\left(x_k^{(i,j)}\right) = \begin{cases} \frac{x_k^{(i,j)} - min_{\mathbf{x}_k}}{max_{\mathbf{x}_k} - min_{\mathbf{x}_k}}, & \text{if } \mathbf{x}_k \text{ is a continuous factor} \\ \left(\mathbf{1}_{\{x_k^{(i,j)}=q_2\}}, \ldots, \mathbf{1}_{\{x_k^{(i,j)}=q_{Q_k-1}\}}\right), & \text{if } \mathbf{x}_k \text{ is a discrete factor} \end{cases}$$

with $p = \sum_{k=1}^{d}(Q_k - 1)$ and $Q_k = 2$ if $\mathbf{x}_k$ is a continuous factor and Q_k is equal to the number of significant levels if $\mathbf{x}_k$ is a discrete factor. Note that $M_{\mathbb{X}}\left(x_k^{(i,j)}\right) \in [0, 1]$. The indicator $I_{\boldsymbol{\beta}}(\mathbb{X}, \mathbf{z})$ varies in $[0, 1]$ and the closer I is to one the larger the set of factors $\mathbb{X}$ is correlated to the response variable. We then have to estimate the set of parameters $\boldsymbol{\beta}$ which maximizes the indicator.

[2]For sake of clarity, we omit here the subscript k.

We propose a two-step optimization procedure realized with a genetic algorithm described in Mebane Jr et al. (2011): first we minimize $\Delta_{M_{\mathbb{X}}\boldsymbol{\beta},\mathbf{z}}$, then for such set of parameters $\boldsymbol{\beta}^*$, we estimate

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^d} r_s^2(M_{\mathbb{X}}\boldsymbol{\beta}, \mathbf{z})(1 + \Delta_{M_{\mathbb{X}}\boldsymbol{\beta}^*,\mathbf{z}}), \tag{6}$$

where $\Delta_{\mathbf{x},\mathbf{y}}$, $\forall(\mathbf{x},\mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n$, is defined by

$$\Delta_{\mathbf{x},\mathbf{y}} = \frac{\sum_{i\in I_0}(R_{x_i}^2 - R_{y_i}^2)}{(n-1)\hat{\sigma}_{R_\mathbf{y}}^2} \tag{7}$$

with $\hat{\sigma}_{R_\mathbf{y}}^2$ is the variance of $R_\mathbf{y}$ and $I_0 = \{i | y_i = 0\}$ (see Appendix A.2 for details).

R codes to implement the methods have been incorporated into the package ZIprop, which is available at `https://gitlab.paca.inrae.fr/meribaud/ziprop`.

4 Simulation studies

In this section, we define a simulated model to investigate the performance of the proposed method.

4.1 Simulated model

The simulated response has to satisfy the constraints described in subsection 2.1. It has to be less than or equal to one (Equation (1)). We consider that the response is exactly equal to one ($\sum_{j=1}^{n_s} z_j^i = 1, \forall j \in \{1,\dots,n_s\}$). However, the simulation method and the result could also be applied to the case "less than". The method used to simulate the response and the factors is described below:

1. Set $n_s > 1$, $n_r > 2$, $m \in \{0.1,\dots,0.3\}$, $d > 1$ and $\boldsymbol{\beta} \in \mathbb{R}^d$.
2. Randomly select $n_0 = \lceil m \times n_s n_r \rceil$ indices in $\{1,\dots,n_s n_r\}$, I_0 gives the set of indices.
3. $\forall(i,j) \in \{1,\dots,n_r\} \times \{1,\dots,n_s\}$ compute the simulated response:

$$\begin{cases} z_j^i \text{ is a realization of the random variable } Y \sim \mathrm{B}(0.1, 0.9) \text{ if } i \notin I_0 \\ z_j^i = 0 \text{ if } i \in I_0 \end{cases}$$

   For a given receiver $i \in \{1,\dots,n_r\}$:
   a. $\exists j \in \{1,\dots,n_s\}$ such that $z_j^i \neq 0$
   b. $\sum_{j=1}^{n_s} z_j^i = 1$.
4. Generate the matrix $\mathbb{X} = (\mathbf{x}_1,\dots,\mathbf{x_d}) \in \mathcal{M}_{n_s n_r \times d}$. $\mathbf{x}_k$, $k \in \{1,\dots,d\}$, is a $n_s n_r$-uplet of the random variable:

$$\begin{cases} X \sim U(0,1) \text{ if } \mathbf{x}_k \text{ is a continuous factor} \\ X \sim \mathbb{U}_{\{0,1\}} \text{ if } \mathbf{x}_k \text{ is a discrete factor.} \end{cases}$$

   where $\text{rank}((\mathbb{X}\beta)^{(i,j)}) = \text{rank}(z_j^i)$ $\forall z_j^i \neq 0$, $(i,j) \in \{1,\dots,n_r\} \times \{1,\dots,n_s\}$.

4.2 Permutation tests and performance indicator

We now test the effect of each factor and compute the performance indicator for different models, setting the number of sources $n_s = 20$, the number of receivers $n_r = 22$, the percentage of non-zeros data $m \in \{0.1, 0.15, 0.2, 0.25\}$ and the number of factors $p = 20$. The first $p/2$ factors are continuous and the last $p/2$ factors are discrete and the vector $\boldsymbol{\beta}$ is defined by:

$$\begin{cases} \beta_k \text{ is a realization of } U(5,10), \text{ if } k = \{1,\dots,5,11,\dots,15\}, \\ \beta_k = 0, \text{ otherwise.} \end{cases}$$

Figure 2 shows the p-values associated with each factor and for different values of m. The figures shows that the m value does not impact the result. The factors $X_{1:5}$ and $F_{1:5}$ are correlated to the response while $X_{6:10}$ and $F_{6:10}$ are not. In most cases the permutation tests have correctly identify the factors i.e. the p-values of $X_{1:5}$ and $F_{1:5}$

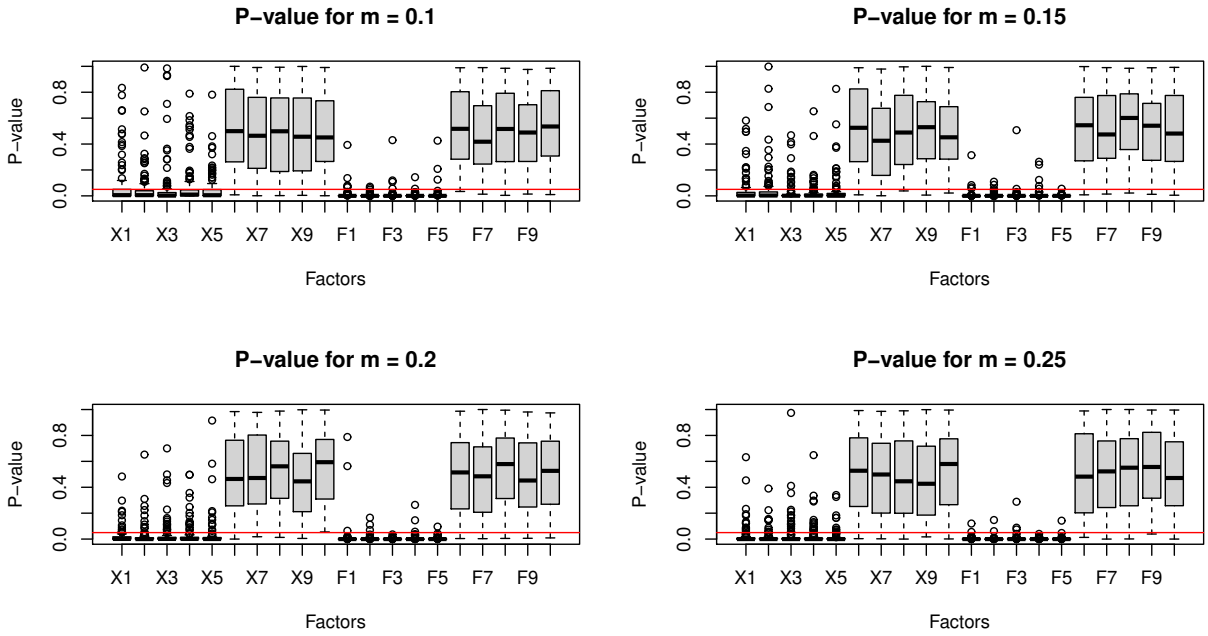

Figure 2: P-values of permutations tests for each factor with $m \in \{0.1, 0.15, 0.2, 0.25\}$. The factors X_k are continuous and F_k are discrete, $k = \{1, \ldots, 10\}$. The data are simulated 100 times for each value of m.

Table 2: Estimated p-value (type I errors) of the permutation tests with 100 repetitions.

m	X6	X7	X8	X9	X10	F6	F7	F8	F9	F10
0.1	0.07	0.05	0.07	0.07	0.06	0.02	0.05	0.03	0.04	0.01
0.15	0.05	0.06	0.02	0.03	0.06	0.04	0.02	0.02	0.05	0.08
0.2	0.05	0.02	0.04	0.03	0.00	0.06	0.06	0.05	0.05	0.04
0.25	0.06	0.07	0.09	0.02	0.08	0.05	0.04	0.03	0.02	0.08

are below the significant level α and the p-values of $X_{6:10}$ and $F_{6:10}$ are above α. The estimated type I errors of the test with different value of m are given in Table 2. The estimated p-values are closed to the significant level $\alpha = 0.05$.

Finally, for each repetition, the performance indicator is computed for the k factors with the lowest p-values, k varying from 2 to 20. We only report the results for $m = 0.2$ as results do not depend on m. The left panel of Figure 3 shows the performance indicator w.r.t the number of significant factors. We can see that the indicator increases until the maximum value around one is reached. The optimal number of selected factors is ten (red line). The indicator is robust in the sense that adding more factors than the optimal number does not affect the performance. This robustness is possible with the estimation of the value of β (see right panel of Figure 3).

In conclusion, our permutation-based approach is a powerful method to identify factors of any type (discrete or continuous) correlated to the response regardless the zero-inflated feature of the data. The indicator is efficient to quantify the part of correlation explained by a set of factors. The value of the indicator increases until all the correlated factor are taking into account in the set. In addition, the indicator is robust to the inclusion of non-correlated factor thanks to the estimation of β. Note that a low estimated value of β does not necessarily imply that the factor is not correlated to the response. That is why permutation tests are strongly recommended.

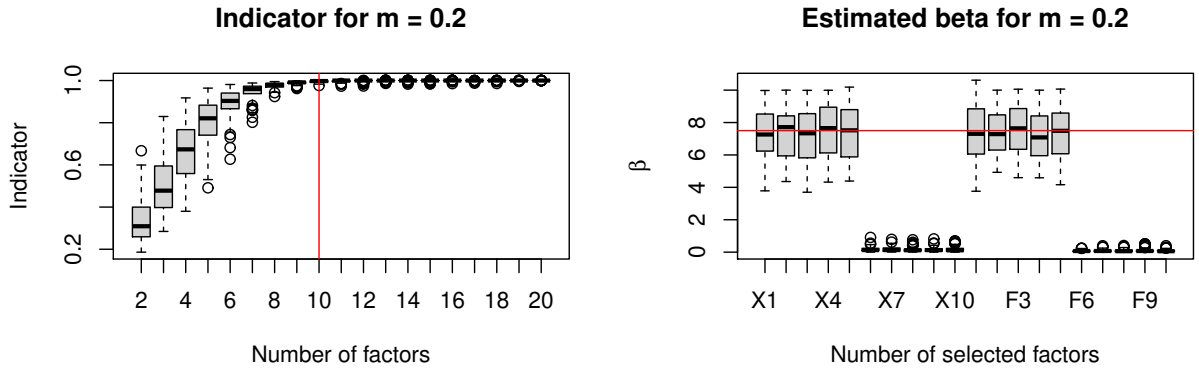

Figure 3: Left: Performance indicator versus the number of factors. Right: Estimated β for each factor. The data are simulated 100 times.

5 Applications

5.1 Equine Influenza

The proportion data for the Equine Influenza in New Market in 2003 represent the transmission links within race horses (hosts). Hughes et al. (2012) inferred the transmission links between 48 hosts: they estimated the probability that each source infects each receiver. In this context, they got many zero probabilities as only a few number of sources may infect each receiver. In addition, the sum of probabilities by receiver is equals to one. We further have factors like age, sex of horses or training yard. The aim is to identify factors that are positively or negatively correlated with these probabilities.

Three factors are qualitative and one is continuous:

1. "Yard": 1 if the receiver and source are trained in the same yard and 0 otherwise,
2. "Sex": "0" if the receiver and source have the same sex "1" otherwise,
3. "Age": 0 if the receiver and source are the same age, 1 for one year difference and $+2$ for more than one year,
4. "distanceYard": distance between training yards for receiver and source.

The factor "Age" has 3 modalities, thus 3 p-values have to be computed. The permutation tests are applied to each factor with 1000 permutations. Results presented in Table 3 show that factors "Yard", "Sex" and "distanceYard" are clearly correlated to the response when "Age" is not.

The statistics of "Yard" and "distanceYard" are negatives, this means that horses trained in the same yard or in a nearby yard have a higher probability of transmitting the disease between them and conversely. This conclusion is really intuitive because horses within training area have more contacts. The statistics of factor "Sex" is also negative and this means that the virus better circulates between horses with the same sex.

Table 3: P-values and statistics associated with permutation tests for each factor.

Factor	p-value	T^*
Yard ("0" - "1")	0	0.033
Sex ("0" - "1")	0.009	0.009
Age	0.637	0.001
distanceYard	0	0.05

Finally, the multivariate analysis is applied to the tree factors "Yard", "Sex" and "distanceYard". The optimal indicator obtained is $I_{\hat{\beta}}(\mathbb{X}, Z) = 0.28$ with $\hat{\boldsymbol{\beta}} = (9.68, 2.21, -8.90)$. This result shows that a weak correlation exists between

these three factors and the transmission links. In addition the two factors related to the yards are certainly correlated.

In conclusion, yards and sex of horses are clearly correlated to the probabilities of transmission. The correlation between yards and transmission is the result of regular contacts between horses that are trained and kept close to each other. The correlation between sex and transmission is more complicated to explain. These factors explain only a little part of the entire correlation and some other factors like groom or transportation would be interesting. Unfortunately, they are not available in this data set.

5.2 Covid-19 (US vs EU)

For the current Covid-19 pandemic, Soubeyrand et al. (2020b) propose a data-driven method based on a mixture model to predict the mortality curve of a focal country using predictive countries. These predictive countries are ahead in terms of death rate compared to the focal country. For any focal country, the mixture model estimates the probability of "following" the same curve than each predictive country. Soubeyrand et al. (2020a) applied this method with US states as focal countries and EU countries as predictive ones. Figure 4 provides an example of such curves and probabilities. They also consider a parametric estimator based on the past of the focal country as a potential predictor. Consequently, the sum of probabilities by focal country is less than one. Lot of zero probabilities are observed because many pairs (US state - EU country) are not similar in terms of curve evolution. The aim is then to disentangle the correlation between

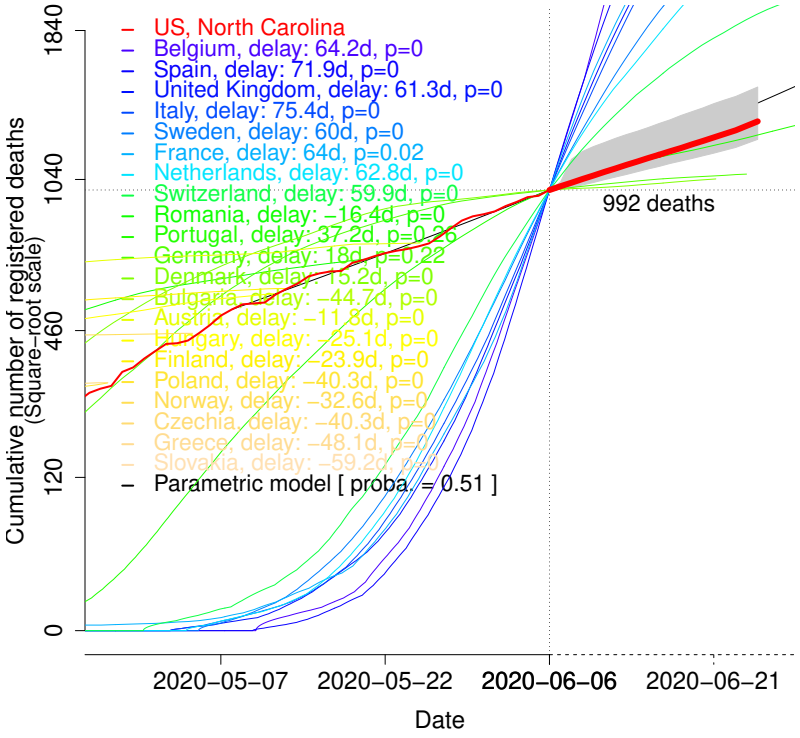

Figure 4: The red line represents the Covid-19 mortality curve of the North Carolina. The selected predictive countries (in green) are Portugal And Germany with respectively a probability of 0.26 and 0.22. The parametric predictor has a probability equals to 0.51. The 6th of June, the North Carolina noticed 992 deaths.

outbreak development and 29 continuous macro factors related to economy, demography, health, healthcare system and climate (see Table 5 in Appendix A.3). The response corresponds to probabilities obtained on June, 6th. In order to construct the data base, the factors related to a specific pair (US state - EU country) are computed from Equation (2).

Figure 5 shows the p-values obtained for each factor. We identify 14 impacting factors: “hospibed”, “smokers”, “lung”, “healthexp”, “urbanpop”, “fertility”, “avghumidity”, “pop_male_65_up”, “gdp_capita”, “physicians_per_1K”, “nurses_per_1K”, “pop_female_0_14”, “gdp2019" and "obesity”. Figure 6 shows the boxplots of the Spearman’s correlation for the fourteen selected factors. Four of them have a positive correlation, that is counter-intuitive. Indeed, the response is the probability that a US curve follows a EU curve, then higher is the probability, more $x^{(i,j)}$ has to be

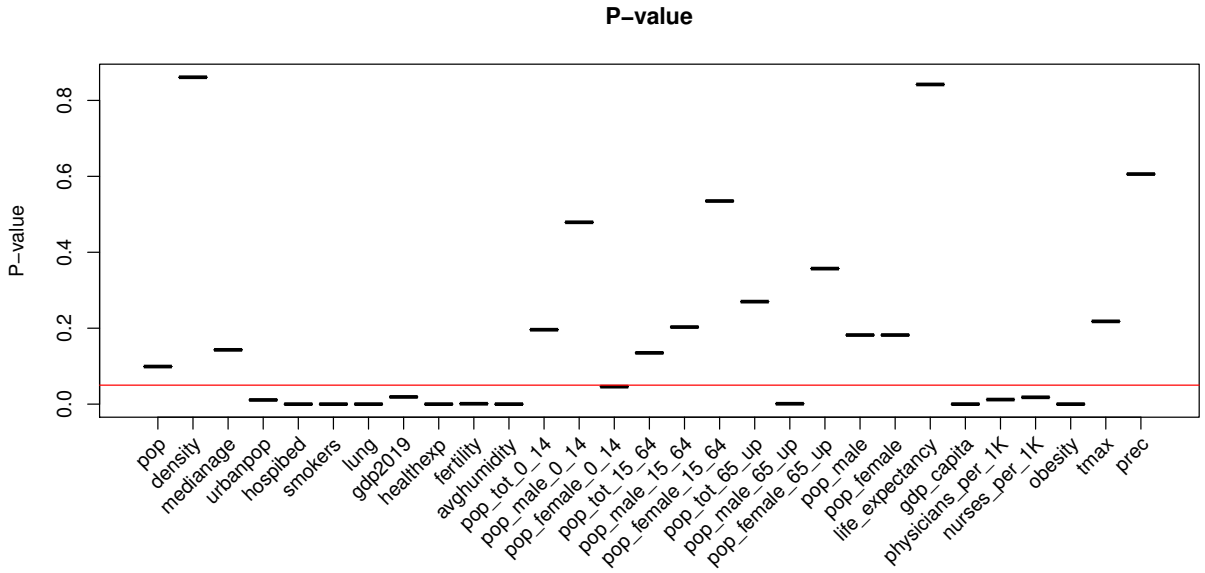

Figure 5: P-value for each factor. The red line is the significance level $\alpha = 0.05$.

low. Consequently, the Spearman's correlation has to be negative. These four factors are "avghumidity", "obesity", "pop_male_65_up" and "physicians_per_1K". This method catches correlations between factors and probabilities, that might not be a causality link. Hence, these negative values might come from indirect links.

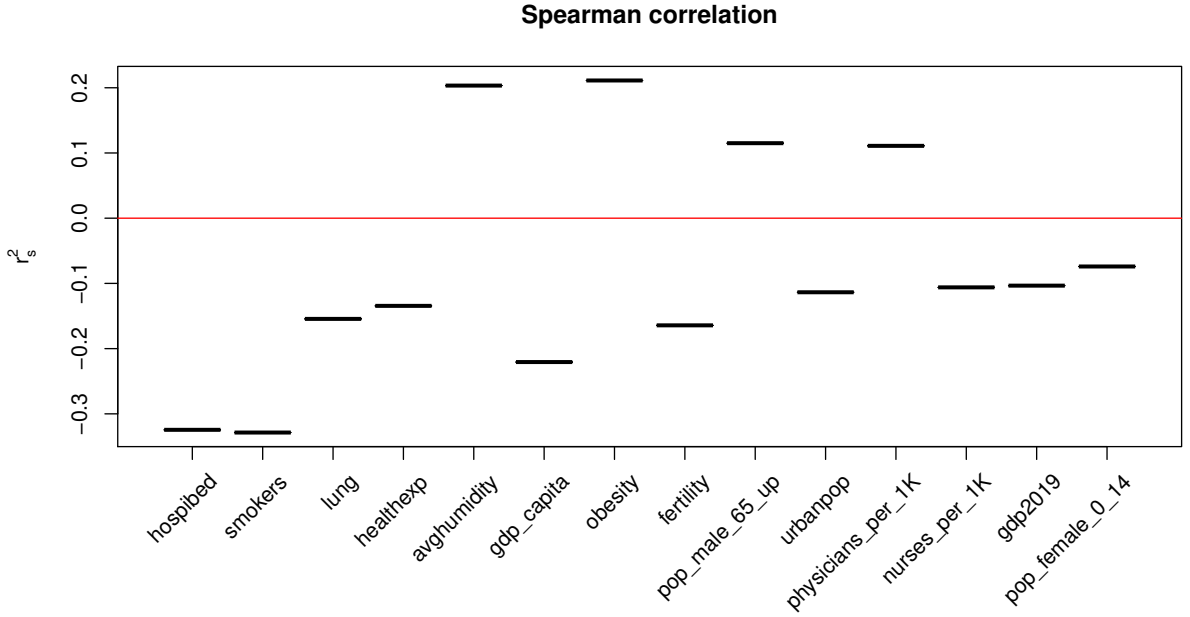

Figure 6: Spearman's correlations calculated on each factor. The red line is $y = 0$.

Then, we applied the multivariate analysis to the ten positively impacting factors. The optimal indicator is $I_{\hat{\boldsymbol{\beta}}}(\mathbb{X}, Z) = 0.71$ (0.81 with all factors, see Appendix A.4 for more details). Estimated parameters $\hat{\boldsymbol{\beta}}$ are given in Table 4. The indicator value shows that a link exists between the probability and these factors.

Table 4: $\hat{\boldsymbol{\beta}}$ corresponding to $I_{\hat{\boldsymbol{\beta}}}(\mathbb{X}, Z) = 0.71$.

Factor	$\hat{\boldsymbol{\beta}}$
"hospibed"	-5.05
"smokers"	-9.94
"lung"	-5.58
"healthexp"	-2.75
"gdp_capita"	-2.06
"fertility"	0.37
"urbanpop"	-0.62
"nurses_per_1K"	-0.34
"gdp2019"	-1.3
"pop_female_0_14"	5.73

Finally, a cross-validation step is realized to ensure the robustness of the methodology. The sample is divided in 100 training (90%) and test (10%) samples. The indicator is optimized on the training sample to obtain the set of parameters $\hat{\boldsymbol{\beta}}_{train}$. The indicator is then computed on the test sample. Figure 7 shows the stability of the indicator optimized in the training sample. The indicators computed from the test samples shows larger variations but remains good in mean. $I_{\hat{\boldsymbol{\beta}}_{train}}(\mathbb{X}_{test}, Z_{test})$.
Figure 8 shows the stability of the estimated parameters. In conclusion, the factors "hospibed", "smokers", "healthexp",

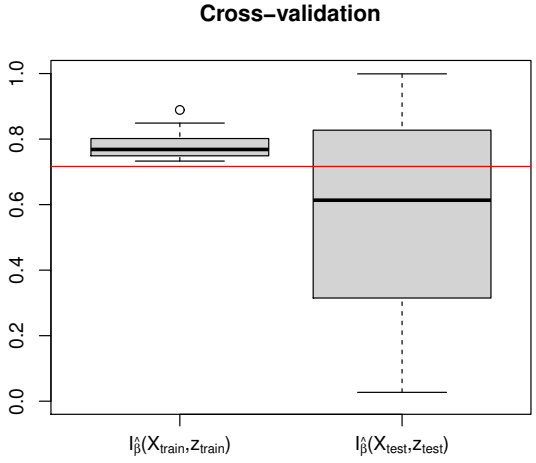

Figure 7: Boxplot of $I_{\hat{\boldsymbol{\beta}}_{train}}(\mathbb{X}_{test}, Z_{test})$. The red line represents $I_{\hat{\boldsymbol{\beta}}}(\mathbb{X}, Z)$ computed from all the observations.

"fertility", "gdp_capita", "lung", "urbanpop" and "nurses_per_1K" impact the similarity measure between US states and EU countries.
In Appendix A.5, the previous cross-validation step is realized with the linear regression and the regression tree. The results show that the two other classical methods are not robust compared with our methodology.

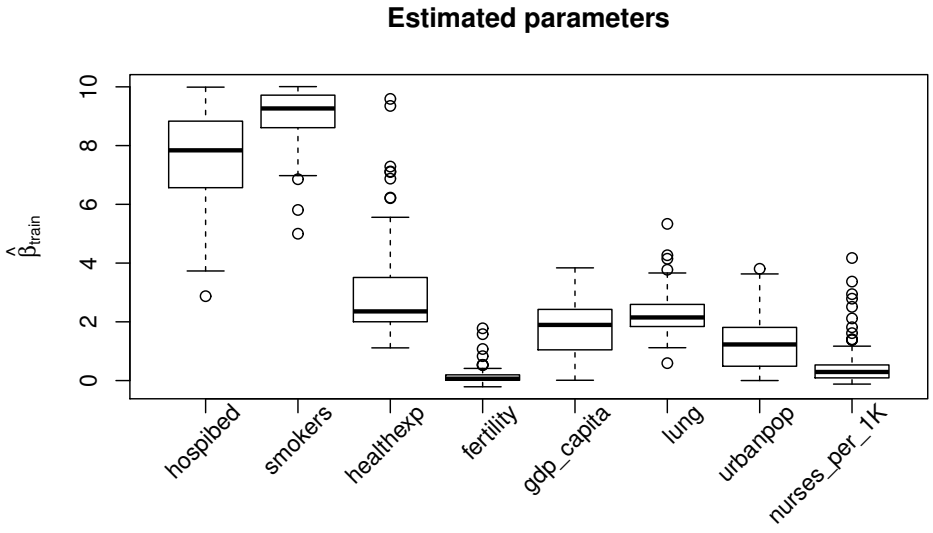

Figure 8: Boxplot of $\hat{\boldsymbol{\beta}}_{train}$ estimated from the test samples.

Appendix

A.1. Why permutations by block?

In this subsection, a degenerate case is presented to show the huge loss of power when classical permutations are done instead of permutations by blocks.
The notations are the same as the ones presented in the paper. Let $n_s n_r$ realizations of a factor X such that $x_1^1 > x_2^1 > \ldots > x_{n_s}^{n_r}$ and $n_s n_r$ realizations of the response Z such that for a fixed receiver i:

$$z_1^i \leq \ldots \leq z_{n_s}^i \tag{8}$$

$$\sum_{j=1}^{n_s} \mathbf{1}_{z_j^i > 0} = c, c \leq n_s \tag{9}$$

$$\sum_{j=1}^{n_s} z_j^i = \frac{i}{n_r} \tag{10}$$

This simulated case can be representative of a real case. For example, in plant epidemics when the spread follows wind gradients, e.g. from East to West and sources and receivers are placed as illustrated in Figure 9. The response are the probabilities of transmission and the factor is the distance between hosts. The closer is a source to a receiver, the higher is the probability of transmission (Equation (8)). Only a given number of hosts are potential sources (Equation (9)). The Equation (10) can comes from an external source that transmits the virus from West to East by another path like underground river. This example is reductive but in Alamil et al. (2019) the authors add a penalization to favor short-distance (geographic or genetic) transmissions.

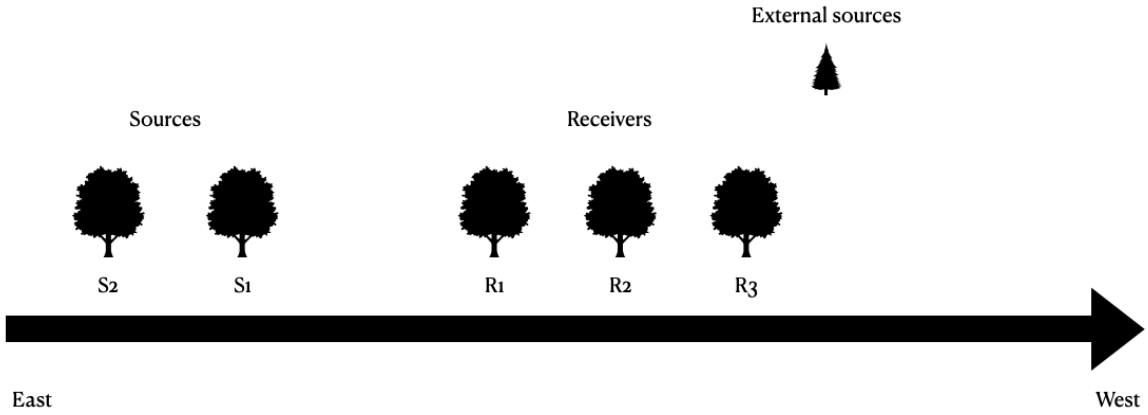

Figure 9: Schematic representation of the position of the trees.

In this context, the factor $\mathbf{x}$ has a huge impact on the response $\mathbf{z}$, then we are under the alternative hypothesis H_1. Let's see how this data set structure impacts the power of permutation tests. Let $1-\beta$ be the power of the test and π a permutation by block of receiver:

$$\begin{aligned} 1-\beta &= 1 - \mathbb{P}(H_0|H_1) \\ &= 1 - \mathbb{P}(T^{\pi} \geq T^*),\ \mathbb{P}(T^{\pi} \geq T^*) = 0 \\ &= 1 \end{aligned}$$

where T is the squared Spearman's correlation. Let π_n a permutation without block constraint:

$$\begin{aligned} 1-\beta =& 1 - \mathbb{P}(H_0|H_1) \\ &= 1 - \mathbb{P}(T^{\pi_n} \geq T^*),\ \mathbb{P}(T^{\pi_n} \geq T^*) >> 0 \\ &<< 1 \end{aligned}$$

In order to illustrate it, let's take $n_r = 10$, $n_s = 20$ and $c = 5$ with 1000 simulated responses. The response is computed as follows, $\forall i \in \{1, \ldots, n_r\}$:

1. Generate c realizations of the random variable $Y \sim \mathcal{U}([0;1])$ written $y_1 \leq \ldots \leq y_c$

2. Compute the simulated response: $(z_1^i, \ldots, z_{n_s}^i) = \frac{i}{n_r \sum_{k=1}^{c} y_k}(0, \ldots, 0, y_c, \ldots, y_1)$

The factor $\mathbf{x}$ is equal to $n_s n_r, n_s n_r - 1, \ldots, 2, 1$.

The estimate power of the permutations tests by blocks is 1 and 0.05 without block (at $\alpha = 0.05$). In conclusion, the permutation by block are crucial to identify factors that are correlated the response variable.

A.2. Calculation of $\Delta_{\mathbf{x},\mathbf{y}}$

This parameter comes from the optimal Spearman's correlation when the rank of two vectors $\mathbf{y^0} \in \mathbb{R}_+^n$ and $\mathbf{x^0} \in \mathbb{R}^n$ are equal except on a given set of indices. In our context, this set correspond to the zeros of the response. Du Bois (1939) gives some formulas for the Spearman's correlation. Kendall (1945) details the calculation of the Spearman's correlation when the vectors $\mathbf{y^0}$ and $\mathbf{x^0}$ have consecutive ties. Here, the elements of calculation are close but it is not exactly the same context.

Let $y_i = R_{y_i^0}$, $x_i = R_{x_i^0}$, $I_0 = \{i | y_i^0 = 0\}$ with $n_0 = \#\{I_0\}$. The rank vectors are assume to be equal $x_i = y_i$ for all $i \notin I_0$. We have $y_i = \frac{n_0+1}{2}$ for all $i \in I_0$ then $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$.
The Spearman's correlation of $\mathbf{y^0}$ and $\mathbf{x^0}$ is equal to the Pearson correlation of $\mathbf{y}$ and $\mathbf{x}$:

$$\begin{aligned}
\hat{r}_s^2(\mathbf{x},\mathbf{y}) &= \hat{r}^2(\mathbf{x},\mathbf{y}) \\
&= \frac{\widehat{Cov}^2(\mathbf{x},\mathbf{y})}{\widehat{\sigma}_{\mathbf{x}}^2 \widehat{\sigma}_{\mathbf{y}}^2}
\end{aligned}$$

$$\begin{aligned}
\widehat{Cov}(\mathbf{x},\mathbf{y}) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})(y_i - \bar{\mathbf{y}}) \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - n\bar{\mathbf{x}}\bar{\mathbf{y}}\right] \\
&= \frac{1}{n-1}\left[y_0 \sum_{i=1}^{n_0} x_i + \sum_{i=n_0+1}^{n} y_i^2 - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n-1}\left[y_0 \sum_{i=1}^{n_0} y_i + \sum_{i=n_0+1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n_0} y_i^2 + \sum_{i=n_0+1}^{n} y_i^2 - n\bar{\mathbf{y}}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\bar{\mathbf{y}}^2\right] \\
&= \widehat{\sigma}_{\mathbf{y}}^2
\end{aligned}$$

$$
\begin{aligned}
x\mathbf{x}^2 &= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{\mathbf{x}}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n_0} x_i^2 + \sum_{i=n_0+1}^{n} y_i^2 - n\bar{\mathbf{y}}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=n_0+1}^{n} y_i^2 + \sum_{i=1}^{n_0} y_i^2 - n\bar{\mathbf{y}}^2 + \sum_{i=1}^{n_0} x_i^2 - \sum_{i=1}^{n_0} y_i^2\right] \\
&= \widehat{\sigma}_{\mathbf{y}}^2 + \frac{1}{n-1}\left[\sum_{i=1}^{n_0}(x_i^2 - y_i^2)\right]
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{\hat{r}_s^2(\mathbf{x},\mathbf{y})} &= \frac{\left(\widehat{\sigma}_{\mathbf{y}}^2 + \frac{1}{n-1}\left[\sum_{i=1}^{n_0}(x_i^2 - y_i^2)\right]\right)\widehat{\sigma}_{\mathbf{y}}^2}{\widehat{\sigma}_{\mathbf{y}}^4} \\
&= \frac{\widehat{\sigma}_{\mathbf{y}}^2\widehat{\sigma}_{\mathbf{y}}^2}{\widehat{\sigma}_{\mathbf{y}}^4} + \frac{\left(\sum_{i=1}^{n_0}(x_i^2 - y_i^2)\right)\widehat{\sigma}_{\mathbf{y}}^2}{(n-1)\widehat{\sigma}_{\mathbf{y}}^4} \\
&= 1 + \frac{\sum_{i=1}^{n_0}(x_i^2 - y_i^2)}{(n-1)\widehat{\sigma}_{\mathbf{y}}^2}
\end{aligned}
$$

$$
\hat{r}_s^2(\mathbf{x},\mathbf{y}) = \frac{1}{1+\Delta_{\mathbf{x},\mathbf{y}}}
$$

where $\Delta_{\mathbf{x},\mathbf{y}} = \frac{\sum_{i=1}^{n_0}(x_i^2 - y_i^2)}{(n-1)\widehat{\sigma}_{\mathbf{y}}^2}$.
Consequently, under the same hypothesis for the vector $y \in \mathbb{R}_+^n$ we have:

$$
\hat{r}_s^2(\mathbf{x},\mathbf{y}) \leq \frac{1}{1+\Delta_{\mathbf{x},\mathbf{y}}} \Leftrightarrow \hat{r}_s^2(\mathbf{x},\mathbf{y})(1+\Delta_{\mathbf{x},\mathbf{y}}) \leq 1
$$

for all vector $x \in \mathbb{R}^n$.

In addition, if $\mathbf{y}$ is such that $y_i \neq y_j$ for all $(i,j) \notin I_0^2$, $i \neq j$ and $\mathbf{x}$ is such that $x_i \neq x_j$ for all $(i,j) \in \{1,\ldots,n\}^2$, $i \neq j$ the parameter $\Delta_{\mathbf{x},\mathbf{y}}$ could be define in a simple way.

$$
\begin{aligned}
\widehat{\sigma}_{\mathbf{y}}^2 &= \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\bar{\mathbf{y}}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n_0}\left(\frac{n_0+1}{2}\right)^2 + \sum_{i=n_0+1}^{n} i^2 - n\left(\frac{n+1}{2}\right)^2\right] \\
&= \frac{1}{n-1}\left[\frac{n_0(n_0+1)^2}{4} + \frac{n(2n+1)(n+1)}{6} - \frac{n_0(2n_0+1)(n_0+1)}{6} - \frac{n(n+1)^2}{4}\right] \\
&= \frac{1}{12(n-1)}\left[n(n+1)(n-1) - n_0(n_0+1)(n_0-1)\right]
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{n_0}(x_i^2 - y_i^2) &= \sum_{i=1}^{n_0} x_i^2 - n_0 y_0^2 \\
&= \sum_{i=n_{x_m}+1}^{n_0} i^2 - \frac{n_0(n_0+1)^2}{4} \\
&= \frac{1}{12}\left[n_0(n_0+1)(n_0-1)\right]
\end{aligned}
$$

$$\begin{aligned}\Delta_{\mathbf{x},\mathbf{y}} &= \frac{n_0(n_0+1)(n_0-1)}{n(n+1)(n-1)-n_0(n_0+1)(n_0-1)} \\ &= \frac{n_0(n_0^2-1)}{n(n^2-1)-n_0(n_0^2-1)}\end{aligned}$$

A.3. Description of factors for Covid-19 application

Table 5: Explanatory factors for the Covid-19 application.

Category	Variable	Description	Unit
Economy	gdp2019	Gross domestic product in 2019	M$
	gdp_capita	Gross domestic product per capita in 2019	$
	healthexp	Health expenditure	M$
Demography	pop	Total population	units
	density	Population density	units per km^2
	urbanpop	Percentage of population living in urban areas	%
	popmale	Percentage of male	%
	pop_tot_0_14	Percentage of population in the age group 0-14 (male, female, total)	%
	pop_tot_15_64	Percentage of population in the age group 15-64 (male, female, total)	%
	pop_tot_65_up	Percentage of population in the age group 65 or more (male, female, total)	%
	mediange	Median age	years
	life_expectancy	Life expectancy at birth	years
Health	lung	Death rate for lung diseases per 100,000 people	units
	fertility	Average number of children per woman	units
	obesity	Percentage of obese people within the popoulation	%
	smokers	Percentage of smokers within the population	%
Healthcare System	hospibed	Number of hospital beds per 1,000 people	units
	physicians_per_1K	Number of physicians per 1,000 people	units
	nurses_per_1K	Number of nurses per 1,000 people	units
Climate	tmin	Average minimum temperature in the first semester	°C
	tmax	Average maximum temperature in the first semester	°C
	prec	Average precipitation in the first semester	mm
	avghumidity	Average relative humidity	%

A.4. Covid-19 indicator with all factors

In this subsection the multivariate analysis is applied to the fourteen factors left, even if the Spearman's correlation is positive. The optimal indicator obtained is $I_{\hat{\boldsymbol{\beta}}}(\mathbb{X}, Z) = 0.81$. The table 6 gives all the value of $\hat{\boldsymbol{\beta}}$ for the selected factors, these values cannot be interpreted because the factors are certainly correlated between them. The indicator value shows that a link exists between the probability and these factors. In section 5.2 the value of the indicator was 0.71, adding these four factors implies a slight increase in the indicator.

Table 6: Value of $\hat{\boldsymbol{\beta}}$ that correspond to $I_{\hat{\boldsymbol{\beta}}}(\mathbb{X}, Z) = 0.81$.

Factor	$\hat{\boldsymbol{\beta}}$
"hospibed"	5.62
"smokers"	9.72
"lung"	7.02
"healthexp"	-3.60
"avghumidity"	-6.07
"gdp_capita"	-1.21
"obesity"	-8.05
"fertility"	1.79
"pop_male_65_up"	1.85
"urbanpop"	6
"physicians_per_1K"	0.18
"nurses_per_1K"	0.27
"gdp2019"	4.41
"pop_female_0_14"	-3.25

A.5. Covid-19 : cross validation to compare multivariate analysis, linear regression and regression tree

In this subsection our methodology is compared to the linear regressions and the decision tree. The factors are selected with the by block-permutation tests procedure. The sample is divided in 100 training (90%) and test (10%) samples. The least square error is minimized on the training set for the additive linear regression. The model is:

$$Z = \boldsymbol{\beta}'\mathbb{X} + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Note that this model assumptions are not satisfied by these data (which are not normality distributed and not independent). However, only the estimated parameters are used. The regression tree learns on the training set with the CART algorithm (Breiman et al. (1984)). For these two models, the R-squared (R^2) is evaluated on the training set and the Q-squared (Q^2) is computed on the test set.

We compared R^2 with $I_{\hat{\boldsymbol{\beta}}_{train}}(\mathbb{X}_{train}, Z_{train})$ on the training set and Q^2 with $I_{\hat{\boldsymbol{\beta}}_{train}}(\mathbb{X}_{test}, Z_{test})$ on the test set. The figure 10 shows that the indicator I remains acceptable when the Q^2 slumps on the test sets. In addition, the indicator is better than the two R^2 evaluated on the training sets. These results confirm that classical models are not appropriate in our context.

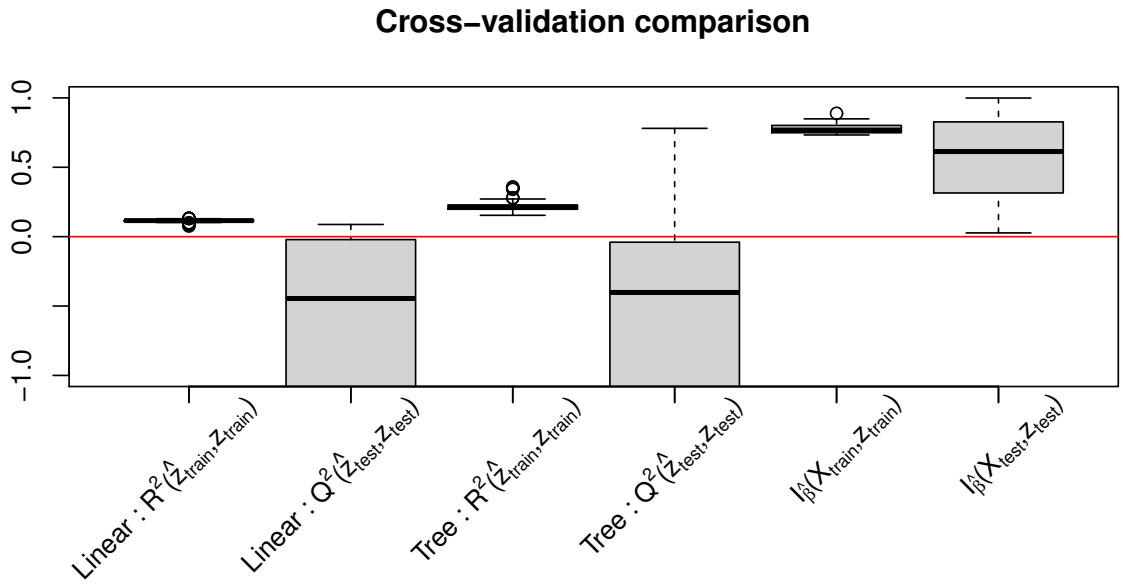

Figure 10: Boxplot of $I_{\hat{\boldsymbol{\beta}}_{train}}(\mathbb{X}_{test}, Z_{test})$ calculated on the 100 test samples. The red line represents $I_{\hat{\boldsymbol{\beta}}}(\mathbb{X}, Z)$ calculated on all the observations.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.

Alamil, M., Hughes, J., Berthier, K., Desbiez, C., Thébaud, G., and Soubeyrand, S. (2019). Inferring epidemiological links from deep sequencing data: a statistical learning approach for human, animal and plant diseases. *Philosophical Transactions of the Royal Society B*, 374(1775):20180258.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis.

Douma, J. C. and Weedon, J. T. (2019). Analysing continuous proportions in ecology and evolution: A practical introduction to beta and dirichlet regression. *Methods in Ecology and Evolution*, 10(9):1412–1430.

Du Bois, P. (1939). Formulas and tables for rank correlation. *The Psychological Record*, 3:46.

Estabrooks, A., Jo, T., and Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20(1):18–36.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hauke, J. and Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93.

Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods*, volume 751. John Wiley & Sons.

Hughes, J., Allen, R. C., Baguelin, M., Hampson, K., Baillie, G. J., Elton, D., Newton, J. R., Kellam, P., Wood, J. L., Holmes, E. C., et al. (2012). Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. *PLoS Pathog*, 8(12):e1003081.

Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika*, pages 239–251.

Mebane Jr, W. R., Sekhon, J. S., et al. (2011). Genetic optimization using derivatives: the rgenoud package for r. *Journal of Statistical Software*, 42(11):1–26.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Pesarin, F. and Salmaso, L. (2010). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons.

Segal, B. D., Braun, T., Elliott, M. R., and Jiang, H. (2018). Fast approximation of small p-values in permutation tests by partitioning the permutations. *Biometrics*, 74(1):196–206.

Shih, J. H. and Fay, M. P. (1999). A class of permutation tests for stratified survival data. *Biometrics*, 55(4):1156–1161.

Soubeyrand, S., Martinetti, D., Ribaud, Mélina Roques, L., and Gabriel, E. (2020a). Comparative analysis of covid-19 outbreaks at the macro level
a comparative analysis of covid-19 outbreaks identifies macro-level indicators of preparedness and vulnerability.

Soubeyrand, S., Ribaud, M., Baudrot, V., Allard, D., Pommeret, D., and Lionel, R. (2020b). The current covid-19 wave will likely be mitigated in the second-line european countries.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Stasinopoulos, D. M., Rigby, R. A., et al. (2007). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7):1–46.

Tang, Z.-Z. and Chen, G. (2019). Zero-inflated generalized dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, 20(4):698–713.

Weisberg, S. (2005). *Applied linear regression*, volume 528. John Wiley & Sons.