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

January 29, 2021

ABSTRACT

The specific nature of zero-inflated proportion data (ZIPD; that are dependent, continuous and bounded) is against using classical supervised methods like linear regression and decision tree to identify impacting factors on a response variable with a ZIPD form. In this article we propose a by-block permutation-based methodology (i) to identify factors (discrete or continuous) that are significantly correlated with ZIPD, and (ii) to define a performance indicator quantifying the percentage of correlation explained by the subset of significant factors. The methodology is illustrated on simulated data and on two real datasets dealing with epidemiology. In the first dataset, ZIPD correspond to estimated probabilities of Influenza transmission within a population of horses. In the second dataset, ZIPD are estimated probabilities that the COVID-19 mortality dynamics in a given geographic entity is similar to those observed in other geographic entities.

1 Introduction

Proportion data are encountered in many fields such as biology, epidemiology and marketing. A common objective is the identification of external factors impacting these data. In marketing, a typical study could be the identification of factors impacting the proportions of products sold to different age groups of customers. In biology, an analogous study could be the analysis of proportions of cures with certain drugs by age groups. In epidemiology, one can be interested in identifying the external factors impacting the transmission of a virus within a host population when the knowledge about the transmissions (i.e., who infected whom) is uncertain (hence, in this case, the probability of transmission from host A to host B can be viewed as a proportion datum). In these typical examples, the 'age groups and products', the 'age groups and drugs' or the 'hosts' can be viewed as the nodes of a network whose edges are weighted by the aforementioned proportions measuring the links between age groups and products / drugs or the links between hosts. The edges from the 'contributing nodes' (i.e., the source hosts transmitting the virus, the products or the drugs) toward a specific 'target node' (i.e., a recipient host or an age group) correspond to a vector of proportions whose sum is equal to one (or eventually lower than one if some contributing nodes are unobserved). Note that in the epidemiological example, recipient hosts can also be source hosts and vice versa (i.e., a host can be both a target and a contributing node).

In this article, we are specifically interested in epidemiological applications. As recently illustrated with the COVID-19 pandemic, grounding strategies for the management of infectious diseases on accurate knowledge about risk factors is paramount for effectively preventing an health crisis. Indeed, assessing the influence of social, biological and environmental factors in the spread of epidemics contributes to identifying levers for controlling the disease dynamics. The spread of epidemics can be approached via the quantification of epidemiological links

between hosts or, more generally, nodes. Typical examples of epidemiological links that we have in mind are: probabilities of disease transmission between individuals (Alamil et al., 2019), and similarity measures of disease dynamics in several geographic entities (Soubeyrand et al., 2020b). Such measures of epidemiological links (i) have an intrinsic correlation structure and (ii) are usually estimated (i.e., uncertain). These features make the investigation of the relationship between epidemiological links and risk factors challenging. Here, we focus on epidemiological links defined as proportions and we aim to provide a statistical methodology reducing the bias of estimation when explaining epidemiological links (hereafter, the response variable) by multiple potentially impacting factors.

Many statistical methods can be used to identify the correlation between factors and a response variable. 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 (e.g., binomial or Poisson), the generalized linear models (GLM) described in Nelder and Wedderburn (1972) can be considered. When one prefers to avoid making a distributional assumption for the response variable, non-parametric predictive models (Hastie et al., 2009) may be a solution. However, non-parametric models do not standardly provide direct testing procedure to identify impacting factors.

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 a usual distribution; see for instance the definition of the zero-inflated Poisson, zero-inflated beta or even zero-inflated binomial distributions in Stasinopoulos et al. (2007). For such zero-inflated models, the influencing factors can also be identified via statistical tests, e.g., in the framework presented by Rigby and Stasinopoulos (2005).

The above-mentioned parametric models are defined for independent and identically distributed (i.i.d.) realizations. However, proportion data are not independent since they sum to a fixed value equal to or lower than one. Such data are often referred to as compositional data, whose mathematical framework is described by Aitchison (1982). Douma and Weedon (2019) propose a classification of compositional data according to the nature of the response (proportions arising from counts *versus* from continuous measurements). Regarding the case of a zero- and/or one-inflated continuous response, the zero- and/or one-inflated beta regression is a solution when the proportions work in pairs (e.g., the proportions of males and females for a given species). When the number of observed categories is greater than two, the Dirichlet's regression can be used. For instance Tang and Chen (2019) propose an adaptation of the zero-inflated Dirichlet regression (ZIDR) model for microbiome compositional data.

Statistical tests are generally associated with the parametric approaches mentioned above for quantifying the significance of a factor (the test generally depends on the type of factors: discrete *versus* continuous). The statistical test accompanying the linear model can treat all types of factors. ANOVA can handle discrete factors with more than 2 levels. The GLM (including zero-inflated data) and the ZIDR can treat continuous factors as well as discrete factors with only 2 levels, even if this restriction is minor since a factor with multiple levels can be treated as several factors with 2 levels each.

In this article, we investigate the relationship between zero-inflated, non-Gaussian, correlated proportion data and several factors of any type. In the epidemiological contexts that we are interested in, this objective translates into the investigation of the impact of individual, environmental, economical, climatic... factors on epidemiological links. Epidemiological links connect pairs of target and contributing nodes, and are measured by inferred probabilities. The structure of the data and the objectives generate constraints on the statistical approach to be used. The response takes values between 0 and 1 (inclusive) and is generally zero-inflated (the zero-inflation makes classical transformations yielding normally-distributed variables inapplicable). Moreover, the realizations are dependent due to the constraint over the sum of probabilities for a given target. Furthermore, factor values for a given target node not only depend on the characteristics of this node but also on the characteristics of the contributing nodes and the target-contributor interaction. Common methods do not match all of these constraints, as illustrated by Table 1. Therefore, we propose a model-free (or more exactly a distribution-free) approach based on permutation tests (Pesarin and Salmaso, 2010) aiming (i) to identify factors (discrete or continuous; characterizing the target, the contributor or the target-contributor pair) that are significant, and (ii) to define a performance indicator quantifying the proportion of correlation explained by the subset of significant factors. We define a by-block permutation test where the permutations are constrained by the dependence structure of data and the test statistic depends on the factor type (the statistic is based on Spearman's correlation if the factor is continuous, and on a difference in mean ranks if it is discrete). The test is applied for each factor separately, but significant factors are then jointly used to compute a performance indicator quantifying the

percentage of correlation explained by the selected set of factors. Figure 1 presents each step of the procedure.

Table 1: Model comparison in their ability to match the constraints considered in this manuscript .

Methods	Response			Factor		Dependency
	Distribution free	$[0, 1]$	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

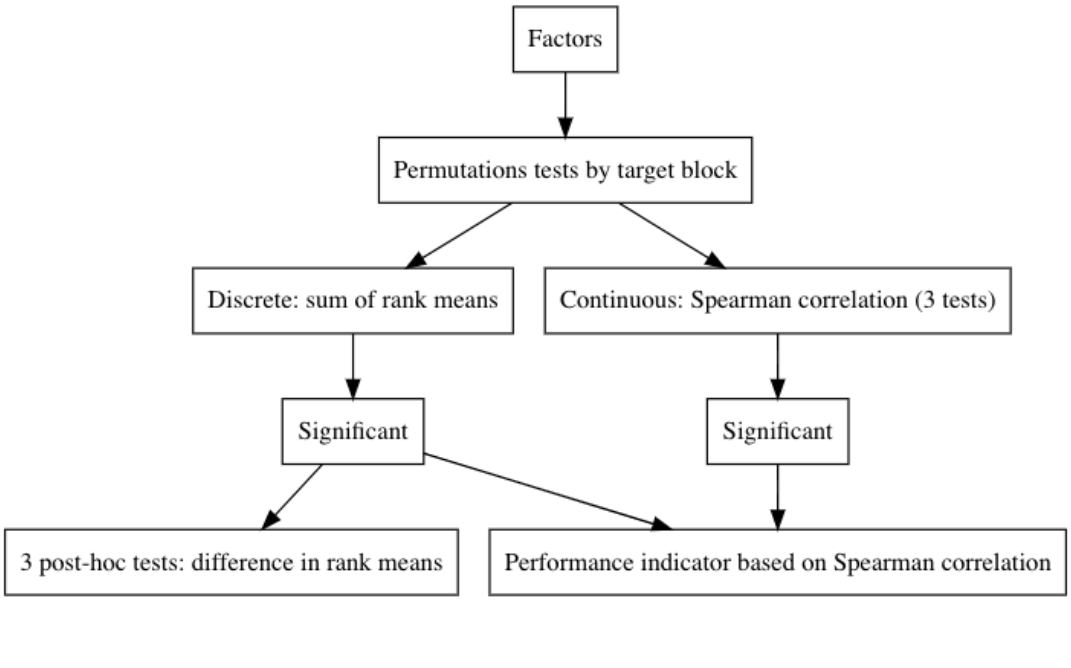

Figure 1: Workflow of the by-block permutation-based methodology.

In what follows, the framework and notations are set in Section 2.1. Section 2 presents the procedure based on permutation tests to identify the factors correlated to the response as well as the performance indicator. Finally, the method is applied to simulations (Section 3) and to real data dealing with Equine Influenza and COVID-19 epidemics (Section 4).

2 Identification and quantification of impacting factors

2.1 Framework and notations

Hereafter, let n_t be the number of target nodes, n_c the number of contributing nodes and d the number of factors.

The response variable Z_j^i is a random variable measuring the (directed) epidemiological link between the target $i \in \{1, \dots, n_t\}$ and the contributor $j \in \{1, \dots, n_c\}$. The higher the value, the stronger the link and the weaker the other links. 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.:

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

The factors characterize any target-contributor pair (i.e., the two nodes and their interaction). We denote by $\mathbb{X} = (\mathbf{x}_1, \dots, \mathbf{x}_d) \in \mathbb{R}^{n_t n_c \times d}$ the set of d factors ($d \in \mathbb{N}^*$). Thus, any pair (i, j) is described by $\left(x_1^{(i,j)}, \dots, x_d^{(i,j)}\right)$.

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:

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

where E is an Euclidean space, $\mathbf{e}^i$ (resp. $\mathbf{e}^j$) is a factor or a set of factors characterizing the target node i (resp. the contributing node j). A classical example of g is the Euclidean distance in $\mathbb{R}^p$ ($p \in \mathbb{N}^*$):

$$\begin{array}{rcl} \|.\|_2 : \mathbb{R}^p \times \mathbb{R}^p & \to & \mathbb{R} \\ (\mathbf{e}^i, \mathbf{e}^j) & \mapsto & x_k^{(i,j)} = \|\mathbf{e}^i - \mathbf{e}^j\|_2. \end{array}$$

If $p = 1$, $\mathbf{e}^i = x_k^i$ and $\mathbf{e}^j = x_k^j$, this distance becomes:

$$\begin{array}{rcl} |.| : \mathbb{R} \times \mathbb{R} & \to & \mathbb{R} \\ (x_k^i, x_k^j) & \mapsto & x_k^{(i,j)} = |x_k^i - x_k^j|. \end{array} \tag{2}$$

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

2.2 A by-block permutation-based approach to identify influencing factors

The specific characteristics of our response variable make the use of classical correlation tests impossible; see e.g. Hollander et al. (2013) for Spearman's test. Indeed, the response has numerous ties (zeros). Kendall (1945) proposes a solution to treat ties in ranking problems. When ties are numerous, some hypotheses on the moments have to be satisfied and checking them may be laborious. Furthermore, the response is dependent within each target-contributors block (see Equation (1)) and classical correlation tests do not take into account such a dependence structure. By-block permutation tests appear to be a possible alternative to take into account these constraints.

Let $\mathbf{x}_k \in \mathbb{R}^{n_t n_c}$, $k = 1, \dots, d$, be the observations of the factor to be tested and let $\mathbf{z} \in \mathbb{R}^{n_t n_c}$ be the observations of the response, whose element (i, j) denoted by z_j^i is the observed value of Z_j^i. We adapt the Conditional Monte Carlo (CMC) algorithm described in Pesarin and Salmaso (2010) for block-permutation to test the correlation between the response and the factor. We denote T the statistic of the test, which depends on the type of factor, and $\lambda_k(\mathbf{z})$ the p-value.

A conditional Monte Carlo algorithm for block-permutation test:

1. Compute the statistic T_k on the original data set $(\mathbf{x}_k, \mathbf{z})$.
2. Do B independent repetitions of: randomly permute the response by block of target nodes, define a new response vector denoted $\mathbf{z}^b$, $b = 1, \dots, B$, and compute the statistic T_k^b on the permuted dataset $(\mathbf{x}_k, \mathbf{z}^b)$.

3. Estimate the p-value by $\hat{\lambda}_k(\mathbf{z}) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\{T_k^b \geq T_k\}}$.

Note that the T_k statistic must be positive to calculate the p-value using the CMC algorithm. In addition, block permutations are required to minimize the second species risk of the test (see Supporting Text S1).

By omitting the subscript k, for a continuous factor $\mathbf{x}$, the statistic T is defined from the non-parametric Spearman's correlation, say $r_s(\mathbf{x}, \mathbf{z})$, between $\mathbf{x}$ and $\mathbf{z}$. The Spearman's correlation is defined as the Pearson's correlation between the rank variables (Spearman, 1904):

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

where ρ is the Pearson correlation, $R_{\mathbf{x}}$ is the random vector that gives the ranks of the elements of $\mathbf{x}$ and $R_{\mathbf{z}}$ is the random vector that gives the ranks of the elements of $\mathbf{z}$. Hence, we define the following tests H_0: "the response and factor ranks are not correlated" versus

(i) H_1: "the response and factor ranks are correlated", and in this case the test statistic is $T = r_s^2(\mathbf{x}, \mathbf{z})$;

(ii) H_1: "the response and factor ranks are positively correlated", and the statistic is $T = r_s(\mathbf{x}, \mathbf{z})$;

(iii) H_1: "the response and factor ranks are negatively correlated", and the statistic is $T = -r_s(\mathbf{x}, \mathbf{z})$.

For a discrete factor $\mathbf{x}$ (still omitting the subscript k) with Q levels, the test hypotheses are H_0: "level-by-level mean ranks are equal" versus H_1: "mean ranks are different for at least two levels" and the statistic corresponds to the one defined in the H-test (Kruskal and Wallis, 1952):

$$T = (n_t n_c - 1)\frac{\sum_{q=1}^{Q} n_q \left(\bar{R}_{\mathbf{z}_{\cdot q}} - \bar{R}_{\mathbf{z}}\right)^2}{\sum_{i=1}^{n_t}\sum_{j=1}^{n_c}(R_{z_j^i} - \bar{R}_{\mathbf{z}})^2}, \tag{3}$$

where $R_{z_j^i}$ denote the rank of the element (i, j) of $\mathbf{z}$, $\bar{R}_{\mathbf{z}} = \frac{1}{n_t n_c}\sum_{i=1}^{n_t}\sum_{j=1}^{n_c} R_{z_j^i}$, $n_q = \sum_{i=1}^{n_t}\sum_{j=1}^{n_c} \mathbf{1}_q(x^{(i,j)})$, $\bar{R}_{\mathbf{z}_{\cdot q}} = \frac{1}{n_q}\sum_{i=1}^{n_t}\sum_{j=1}^{n_c} R_{z_j^i}\mathbf{1}_q(x^{(i,j)})$ and $\mathbf{1}_q(x^{(i,j)}) = 1$ if $x^{(i,j)} = q$, $\mathbf{1}_q(x^{(i,j)}) = 0$ otherwise.

If the p-value is low enough such that the factor is considered as significant (i.e., below the significance level), "post-hoc" tests can be constructed to test the impact of factor levels. Let q and $\tilde{q}$ be two levels, we can then make the following tests H_0: "there is no difference between the two mean ranks" versus

(i) H_1: "there is a difference between the two mean ranks" and the statistic is $T = (\bar{R}_{z_{\cdot q}} - \bar{R}_{z_{\cdot \tilde{q}}})^2$;

(ii) H_1: "the mean ranks of the level q is lower than the mean ranks of $\tilde{q}$" and the statistic is $T = \bar{R}_{z_{\cdot \tilde{q}}} - \bar{R}_{z_{\cdot q}}$;

(iii) H_1: "the mean ranks of the level q is greater than the mean ranks of $\tilde{q}$" and the statistic is $T = \bar{R}_{z_{\cdot q}} - \bar{R}_{z_{\cdot \tilde{q}}}$.

Note that the statistics are the difference in mean ranks defined by Dunn (1964). In addition, if the discrete factor has more than two levels, the problem becomes a multiple comparison problem. A correction can be applied accordingly to control the occurrence of false positives, e.g., the Bonferroni correction which consists in multiplying the p-values by the number of comparisons, or the less conservative 'improved Bonferroni correction' introduced by Hochberg (1988). As an illustration, we provide the basic and the Hochberg-corrected p-values for the post-hoc tests performed in the application dealing with equine influenza.

2.3 A performance indicator to quantify the monotonous dependency

The previous section presents an approach to identify factors that are individually correlated to the response. However, the multivariate aspect of the correlation is not taken into account. Here we deal with it by developing a performance indicator that simultaneously takes into account all discrete and continuous factors previously identified. The set of factors is represented by a single linear combination of all factors. We set the linear combination as in a regression without the intercept, which has no impact since the indicator is based on ordering. The indicator is defined as the ratio between the Spearman correlation and its upper bound. This bound represents the maximum Spearman correlation that can be obtained with the given set of factors by taking into account the structure of the response (zero inflation and ties). The indicator can be viewed as a surrogate for the coefficient of determination used in linear regression. It represents the monotonous relationship between the combination of factors and the response while taking into account the particular structure of the response. The closer the indicator to 1, the stronger the relationship.

Thus the expression of the performance indicator satisfies:

$$I_{\boldsymbol{\beta}}(\mathbb{X}, \mathbf{z}) = r_s^2(M_{\mathbb{X}}\boldsymbol{\beta}, \mathbf{z})(1 + \Delta_{M_{\mathbb{X}}\boldsymbol{\beta}, \mathbf{z}}), \tag{4}$$

where the upper bound is $\frac{1}{1+\Delta_{M_{\mathbb{X}}\boldsymbol{\beta},\mathbf{z}}}$ and the elements of the design matrix $M_{\mathbb{X}} \in \mathbb{R}^{n_t n_c \times d'}$ are defined by:

$$M_{\mathbb{X}}(\ell,k) = \begin{cases} \dfrac{x_k^{(i,j)} - \min\{\mathbf{x}_k\}}{\max\{\mathbf{x}_k\} - \min\{\mathbf{x}_k\}}, & \text{if } \mathbf{x}_k \text{ is a continuous factor} \\ \left(\mathbf{1}_q(x_k^{(i,j)})\right)_{q=1,\ldots,Q_k}, & \text{if } \mathbf{x}_k \text{ is a discrete factor,} \end{cases}$$

$\ell = (i-1)n_c + j$ represents the ℓ-th row and k the k-th column of the design matrix $M_{\mathbb{X}}$, $\min\{\mathbf{x}_k\}$ (resp. $\max\{\mathbf{x}_k\}$) is the minimum (resp. maximum) element of the vector $\mathbf{x}_k$, $d' = \sum_{k=1}^{d} Q_k$, $Q_k = 1$ if $\mathbf{x}_k$ is a continuous factor and Q_k is equal to the number of levels if $\mathbf{x}_k$ is a discrete factor. Note that $M_{\mathbb{X}}(\ell,k) \in [0,1]$. The indicator $I_{\boldsymbol{\beta}}(\mathbb{X},\mathbf{z})$ varies in $[0,1]$; the larger $I_{\boldsymbol{\beta}}(\mathbb{X},\mathbf{z})$, the larger the correlation between the set of factors $\mathbb{X}$ and the response variable. We then have to estimate the set of parameters $\boldsymbol{\beta}$ which maximizes the indicator. The values of the components of $\boldsymbol{\beta}$ associated with the factors identified as insignificant are set to zero, and the optimization is carried out with respect to the remaining subset of parameters (of dimension $d'' \leq d'$) using the genetic algorithm described by Mebane Jr et al. (2011). Hence, we estimate $\boldsymbol{\beta}$ as follows:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\mathbb{R}^{d''}} r_s^2(M_{\mathbb{X}}\boldsymbol{\beta},\mathbf{z}), \tag{5}$$

and we calculate the performance indicator by pluging-in $\hat{\boldsymbol{\beta}}$:

$$I_{\hat{\boldsymbol{\beta}}}(\mathbb{X},\mathbf{z}) = r_s^2(M_{\mathbb{X}}\hat{\boldsymbol{\beta}},\mathbf{z})(1+\Delta_{M_{\mathbb{X}}\hat{\boldsymbol{\beta}},\mathbf{z}}), \tag{6}$$

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 $\hat{\sigma}_{R_{\mathbf{y}}}^2$ the variance of $R_{\mathbf{y}}$ and $I_0 = \{i | y_i = 0\}$ (see Supporting Text S2 for details).

The solution of the maximization (5) is obviously not unique (if $\boldsymbol{\beta}_0$ is a solution, $a\boldsymbol{\beta}_0$ is also a solution for all real value $a \neq 0$), but this is not an issue in the proposed framework since only the rank are taken into account and $I_{\boldsymbol{\beta}_0}(\mathbb{X},\mathbf{z}) = I_{a\boldsymbol{\beta}_0}(\mathbb{X},\mathbf{z}), \forall a \neq 0$.

2.4 Rank prediction

The approach presented in this paper can give a prediction of the ranks of known or unknown contributors for unknown targets. The factors used to estimate the performance indicator must be available for the new individuals studied. For example, for different targets and a set of potential contributors, the predicted rank vector is given by $\hat{R}_{\mathbf{z}} = M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$. The matrix $M_{\mathbb{X}_i}$ is computed on the set of selected factors for the selected target contributor pairs.

Rank prediction can also be performed using classical prediction methods such as linear regression or regression trees. The models give a prediction of the response. The ranks are calculated based on this prediction. However, these methods do not work directly on the ranks. The article of Kloke and McKean (2012) proposes a R package allowing estimation based on ranks for linear models. In this context, the models are built using the factors identified by the tests presented in the subsection 2.2. The difference between the linear model and the rank-based linear model lies in the method of parameter estimation.

To investigate the robustness and the quality of the performance indicator, we compared our multivariate analysis (MA) with the linear regression model (LM), the linear regression model based on rank (LMRank) and the decision tree (Tree) using cross-validation. Target hosts are randomly divided into a train sample with 80% of the targets and a test sample with 20% of the targets. The tests for selecting the factors are applied to the global sample (union of train and test samples), and the performance indicator as well as the contributor ranking indicator (CR) are computed from both the train sample and the test sample. This procedure is independently repeated 100 times. The CR indicator is the average over the targets of 'the proportion of the N_j^i contributors with positive transmission probabilities for target i that are ranked in the top N_j^i contributors by the predictor under consideration (MA, LM, LMRank or Tree)'; see Supporting Text S5.

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

3 Simulation study

In this section, we carry out a simulation study to investigate the performance of the proposed method.

3.1 Simulated data

The simulated response has to satisfy the constraints described in Section 2.1. Hence, we simulate data accordingly and, for the constraint over the sum, we consider that it is exactly equal to one. The algorithm applied to simulate the response and the factors is described below:

1. Set values for $n_c > 1$, $n_t > 2$, $m \in [1/n_c, 1]$ (the proportion of non-negative values for the responses z_j^i), $d > 1$ and $\boldsymbol{\beta} \in \mathbb{R}^{d'}$.
2. Randomly and uniformly draw $n_0 = \lceil (1-m) n_c n_t \rceil$ indices in $\{1, \ldots, n_t n_c\}$ (where $\lceil \cdot \rceil$ is the ceiling function); I_0 gives the set of drawn indices corresponding to responses equal to zero.
3. Generate the response vector $\mathbf{z} \in \mathbb{R}^{n_t n_c}$ such that its elements z_j^i are independently drawn from the beta distribution $\mathrm{B}(0.1, 0.9)$ if $i \notin I_0$ or set at zero if $i \in I_0$ and then scaled as follows for each target $i \in \{1, \ldots, n_t\}$: $z_j^i \leftarrow \frac{z_j^i}{\sum_{j=1}^{n_c} z_j^i}$. Therefore, the response is simulated in such a way that for a given target $i \in \{1, \ldots, n_t\}$:
   a. $\exists j \in \{1, \ldots, n_c\}$ such that $z_j^i > 0$,
   b. $\sum_{j=1}^{n_c} z_j^i = 1$.
4. Generate the matrix $\mathbb{X} = (\mathbf{x_1}, \ldots, \mathbf{x_d}) \in \mathbb{R}^{n_t n_c \times d}$ such that the components of the $n_t n_c$-uplet $\mathbf{x}_k$, $k \in \{1, \ldots, d\}$, are independently drawn from the uniform distribution $U(0,1)$ over the interval $[0,1]$ if $\mathbf{x}_k$ is a continuous factor and from the uniform distribution $U(\{0,1\})$ over the set of two values $\{0,1\}$ if $\mathbf{x}_k$ is a discrete factor, and such that

$$\text{rank}\left((\mathbb{X}\boldsymbol{\beta})^{(i,j)}\right) = \text{rank}(z_j^i), \quad \forall z_j^i \neq 0,\ \forall (i,j) \in \{1, \ldots, n_t\} \times \{1, \ldots, n_c\},$$

   where $(\mathbb{X}\boldsymbol{\beta})^{(i,j)}$ is the term (i,j) of $\mathbb{X}\boldsymbol{\beta}$.

3.2 Simulation specification

We test the effect of each factor and compute the performance indicator for different models, setting the number of contributing nodes $n_c = 20$, the number of target nodes $n_t = 22$, the proportion of non-zero data $m \in \{0.1, 0.15, 0.2, 0.25\}$ and the number of factors $d = 20$. The first $d/2$ factors are continuous, the last $d/2$ factors are discrete and the vector $\boldsymbol{\beta}$ satisfies:

$$\begin{aligned}\boldsymbol{\beta} = (&\beta_1, \beta_2, \beta_3, -\beta_4, -\beta_5, 0, 0, 0, 0, 0,\\ &\beta_1, 0, \beta_2, 0, \beta_3, 0, 0, \beta_4, 0, \beta_5, \beta_5, \beta_5, 0, 0, 0, 0, 0, 0, 0, 0)\end{aligned}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{3d/2}$ and β_k, $k = \{1, \ldots, 5\}$, are independently drawn from the uniform distribution $U(5, 10)$. The first 10 components of β correspond to the continuous factors, the 10 following pairs of components of β correspond to the discrete factors (each factor having two modalities).

3.3 Estimated errors of permutation tests

Here, we first assess the performance of the two-tailed permutation test for continuous and discrete factors and different proportions of non-zero data. Figure 2 shows the distribution of p-values for each factor and each value of m. The factors $\mathbb{X}_{1:5}$ and $\mathbb{X}_{11:15}$ are generally identified as correlated to the response while $\mathbb{X}_{6:10}$ and $\mathbb{X}_{16:20}$ are not (i.e., the p-values of $\mathbb{X}_{1:5}$ and $\mathbb{X}_{11:15}$ are generally below the significance level $\alpha = 0.05$ whereas the p-values of $\mathbb{X}_{6:10}$ and $\mathbb{X}_{16:20}$ are generally above α). The estimated type I errors of the test at the risk level 0.05 are given in Table 2 for different values of m, and show that the test is relatively well calibrated for the diverse configurations that are considered. The type II errors provided by Table 3 are very small for discrete factors whatever the value of m. In contrast, the type II errors for continuous factors are larger (0.12 in average) and decrease with m. Hence, the power of the test is very large for discrete factors and is correct for continuous factors (with values of β that we consider).

We carried out the same analysis for the one-tailed permutation tests. Very similar results are obtained as shown by Supporting Tables S1–S4.

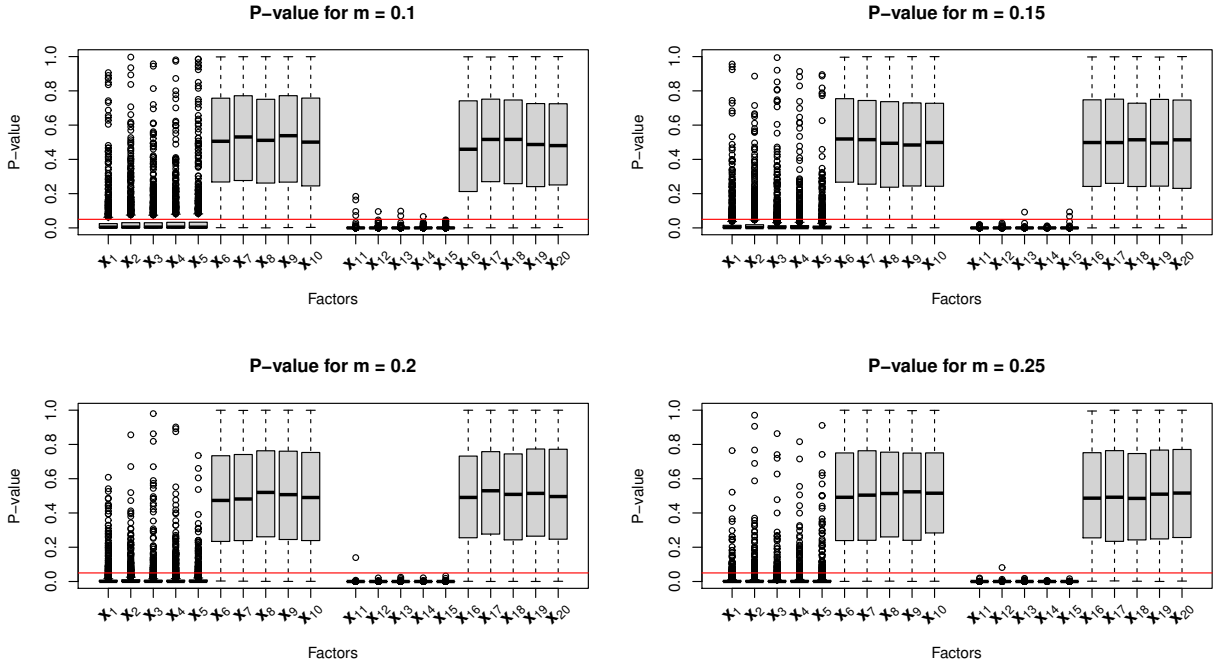

Figure 2: P-values of two-tailed permutation tests for each factor with $m \in \{0.1, 0.15, 0.2, 0.25\}$. The factors $\mathbf{x}_k$ are continuous for $k = \{1, \ldots, 10\}$ and discrete for $k = \{11, \ldots, 20\}$. The data are simulated 1000 times for each value of m.

Table 2: Estimated type I errors of the two-tailed permutation tests with 1000 repetitions.

	Continuous factors					Discrete factors				
m	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$	$\mathbf{x}_{16}$	$\mathbf{x}_{17}$	$\mathbf{x}_{18}$	$\mathbf{x}_{19}$	$\mathbf{x}_{20}$
0.1	0.055	0.044	0.049	0.033	0.045	0.057	0.041	0.048	0.049	0.053
0.15	0.041	0.048	0.048	0.040	0.054	0.059	0.046	0.045	0.043	0.050
0.2	0.059	0.049	0.051	0.060	0.049	0.054	0.040	0.051	0.063	0.054
0.25	0.050	0.053	0.062	0.042	0.046	0.054	0.039	0.045	0.038	0.055

Table 3: Estimated type II errors of the two-tailed permutation tests with 1000 repetitions.

	Continuous factors					Discrete factors				
m	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$	$\mathbf{x}_{14}$	$\mathbf{x}_{15}$
0.1	0.174	0.194	0.189	0.190	0.205	0.004	0.001	0.002	0.001	0.000
0.15	0.128	0.157	0.127	0.112	0.110	0.000	0.000	0.001	0.000	0.002
0.2	0.090	0.093	0.089	0.095	0.091	0.001	0.000	0.000	0.000	0.000
0.25	0.064	0.066	0.054	0.075	0.076	0.000	0.001	0.000	0.000	0.000

3.4 Performance indicator

For each repetition performed for $m = 0.25$ (yielding the largest test power for continuous variables), the performance indicator is computed for the k factors with the lowest p-values, k varying from 2 to 20. Figure 3 (left) shows the distribution of the performance indicator with respect to k. The indicator increases until it reaches a plateau at the value one (which is the maximum value of the indicator) approximately when $k = 10$ (red line), which corresponds to the actual number of factors having a significant effect. The indicator is robust in the sense that adding more factors than the actual number of factors with significant effects does not affect the performance. This robustness is consistent with the adequate estimation of β. Indeed, Figure 3 (right) shows that estimated coefficients for insignificant factors

are close to zero, while estimated coefficients for significant factors take values between approximately 5 and 10 (or -10 and -5), i.e., the range of actual values of $\beta_1, \ldots, \beta_5$. During the optimization procedure the range of variation of the parameters $\boldsymbol{\beta}$ is $[-10; 10]$. It is possible to widen this range of variation and in this case the estimated values will be pushed towards the limits of the range. The vector of estimated parameters on the expanded domain $\mathcal{D}_e$ will be approximately equal to a constant to the vector of the restricted domain $\mathcal{D}_r$ i.e. $\hat{\boldsymbol{\beta}}_{\mathcal{D}_e} = c \times \hat{\boldsymbol{\beta}}_{\mathcal{D}_r}$, $c \in \mathbb{R}$. Therefore, the value of the estimator will remain the same.

The main conclusion of this simulation study is that the by-block 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 performance indicator is efficient to quantify the monotonous dependence between the set of factors and the response. The value of the indicator increases until all the correlated factors are taken into account in the set of explaining factors. In addition, the indicator is robust to the inclusion of non-correlated factor thanks to the adequate estimation of β.

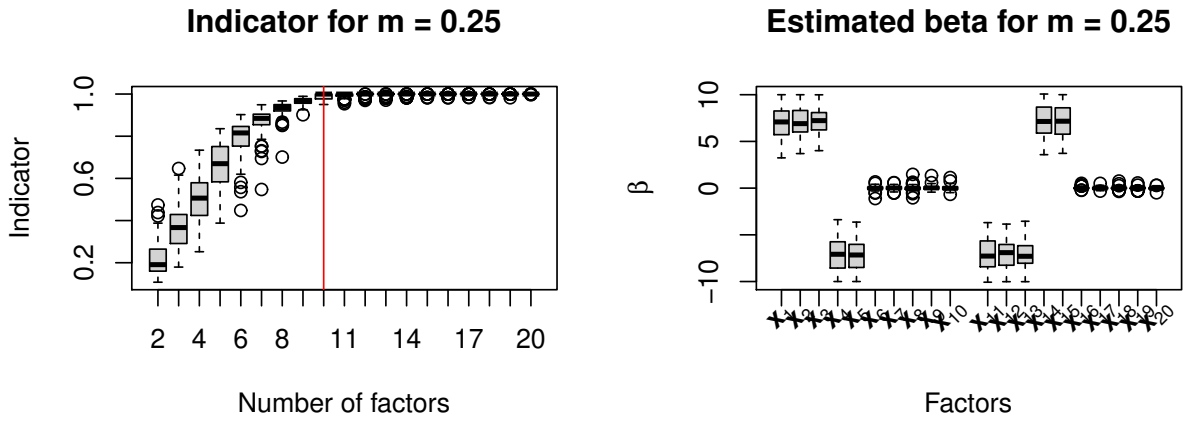

Figure 3: Left: Distribution of the performance indicator for varying number of included factors (factors are successively incorporated by first including those with lowest p-values). Right: Distribution of estimated coefficients (appearing in β) for each factor. The data are simulated 100 times.

3.5 Cross validation

In this sub-section, we will compare by cross-validation the methodology proposed in this article (MA) with the three other methods introduced in the sub-section 2.4 (LM, LMRank and Tree). The prediction methods are computed on the set of factors selected by the permutation tests.

Figure 4 and S6 shows the good performance of the multivariate analysis (MA). The linear models (LM and LMRank) gives good results too. This result can be explained by the way the simulations were constructed. Indeed, we assume that the ranks of the response are equal to the ranks of a linear combination of the factors. The decision tree (Tree) gives poor results in this context. In addition, the two indicators remains stable along train samples and test samples. We observe that the two indicators have quite similar values.

4 Applications

4.1 Equine Influenza

Here we consider the Equine Influenza outbreak in New Market in 2003 throughout a population of race horses distributed in several yards. Genomic data collected during this outbreak from 48 hosts were presented and studied by Hughes et al. (2012) to explore the virus transmissions across the observed horse population. These data and the BadTrIP software (De Maio et al., 2018) were used to jointly estimate the probabilities of the disease transmission

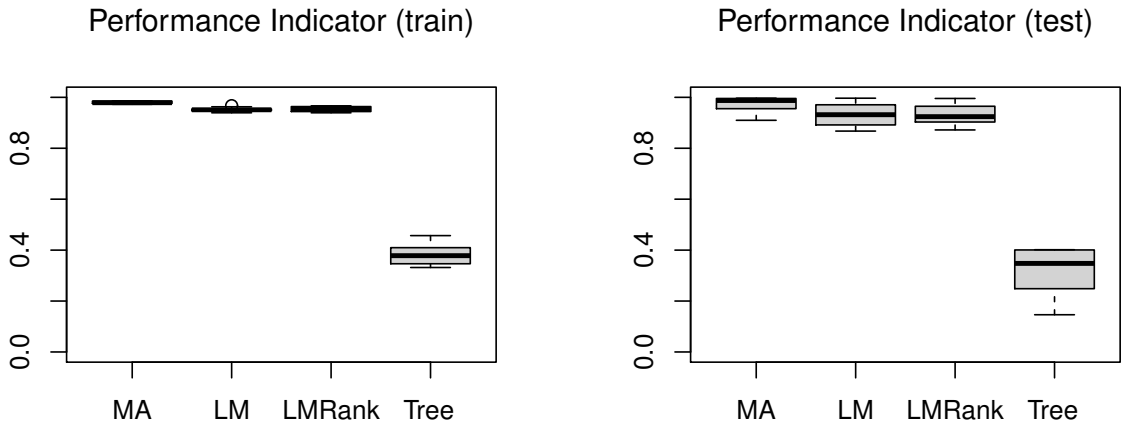

Figure 4: Boxplots of the performance indicator calculated from the train and test samples for MA, LM, LMRank and Tree in the simulations.

between infectious and susceptible hosts (see also Alamil, 2020, chap. 3). The estimated transmission probabilities (shown in Supporting Figure S2) are used in the present study as response proportion data. Many of the estimated transmission probabilities are equal to zero, which means that, for each target, BadTrIP identified only a small number of potential contributors. Moreover, we use variables related to age, sex and training yard as potentially explaining factors. We considered four discrete factors and one continuous factor:

1. "Same_Yard": 1 if the target and contributor are trained in the same yard, 0 otherwise;
2. "Same_Sex": 1 if the target and contributor have the same sex, 0 otherwise;
3. "Diff_Age": 0 if the target and contributor have the same age, 1 for a one-year difference and 2 for more than one year;
4. "Dist_Yard": geographic distance (in km) between the training yards of the target and the contributor;
5. "Trans_Sex": "F→F" if a female infected another female, "M→F" if a male infected a female, "F→M" if a female infected a male and "M→M" if a male infected another male.

Some of these factors are missing for some source-receptor pairs (see Supporting Table S5). Hence, the tests for assessing the effect of a given factor on the transmission probability are applied on the subset of complete data for this factor.

Factors "Same_Yard", "Same_Sex", "Dist_Yard" and "Trans_Sex" are significantly correlated to the transmission probability whereas "Diff_Age" is not; see Table 4. The test statistics of "Same_Yard" and "Dist_Yard" being negative, horses trained in the same yard or in nearby yards have a higher chance to be linked by a transmission. This is a clearly intuitive result certainly due to higher contact rate in shared training areas. The statistics of post-hoc univariate tests for factor "Same_Sex" is also negative, which means that the virus better circulates between horses with the same sex. Moreover, the post-hoc tests on the "Trans_Sex" modalities show that only the difference between "F→M - M→M" (and "M→F - M→M" when one considers the corrected p-values) are not significant. The results on the p-values and the sign of the statistics show that transmissions between females are favored compared to all other possible combinations (F→F transmissions have positive probabilities 1.8 times more than expected under complete randomness; see Supporting Table S6). In addition, there is more intersex transmission when females are the sources. Supporting Text S4 shows that the significance of gender-related factors is neither confounded with the effect of the other factors available in the data set nor a consequence of heterogeneous sex frequencies.

The calculation of the performance indicator leads to the value 0.21 using the four selected factors. This relatively low value, which indicates that there is a moderate correlation between the combination of the four factors and the transmissions, can actually be viewed as quite large given the fact that we only consider very basic factors to *predict* the transmissions.

Table 4: Test results for the equine influenza application. Top: Statistic (T), p-value (pv) and Spearman's correlation (r_s; for the continuous factor only) associated with the two-sided permutation tests performed for the five factors. Bottom: Statistic (T), p-value (pv), Hochberg-corrected p-value (pv*; for discrete factors with more than 2 levels) associated with the post-hoc permutation tests applied to significant discrete factors. Lines with a significant p-value are highlighted in gray.

Factor	T	pv	r_s
Same_Yard	0.05	0	
Same_Sex	0.007	0.031	
Diff_Age	0.001	0.8	
Dist_Yard	0.05	0	-0.22
Trans_Sex	0.042	0	

Factor	Factor level	T	pv	pv*
Same_Yard	0 - 1	-444	0	
Same_Sex	0 - 1	-24.77	0.04	
Trans_Sex	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

To investigate the robustness and the quality of the performance indicator, we compared our multivariate analysis by cross-validation, see sub-section 2.4. Figure 5 and Figure S7 show that the three methods are relatively robust in the sense that the indicators take similar average values in the train and test samples. The decision tree slightly outperform the two other methods in terms of prediction measured by the performance indicator, possibly because most of the factors are discrete (3 over 4) and the only continuous factor ("Dist_Yard") takes only 37 different values out of 650 observations. When one measures the prediction ability with the CR indicator, our multivariate analysis and the decision tree have similar performance.

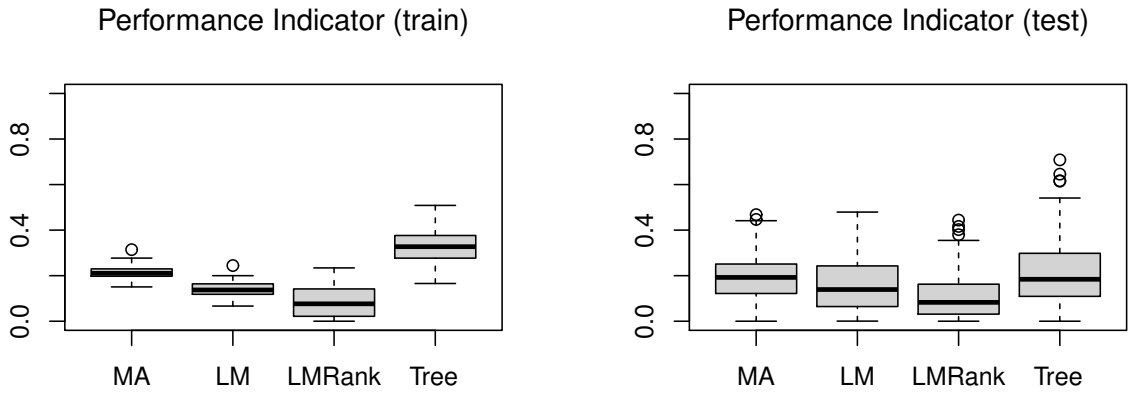

Figure 5: Boxplots of the performance indicator calculated from the train and test samples for MA, LM and Tree in the equine influenza study.

4.2 COVID-19

Soubeyrand et al. (2020b) proposed a data-driven method to predict the mortality curve of a target country with a mixture of the mortality curves of countries that are ahead of time in terms of mortality rate. The mixture is more exactly formed by the mortality curves of *contributing countries* as well as an additional parametric predictor, and the method is essentially grounded on the estimation of the mixture probabilities. Real-time predictions based on this method are available for more than 100 countries via the following web application: `http://covid19-forecast.biosp.org/`.

Here, we use the estimated mixture probabilities as proportion data. Targets are states from the USA and provinces from Canada; contributors are members of the European Economic Area (EEA) and the European Free Trade Association (EFTA). We only consider geographic entities with at least 5,000,000 inhabitants (leading to 25 targets and 21 contributors) and the first epidemic wave by using data up to June 6, 2020. Mortality data used to estimate the mixture probabilities were collected from the Johns Hopkins University Center For Systems Science and Engineering (**?**) and The Covid Tracking Project (`https://covidtracking.com`). The choice of considering Northern American targets and European contributors was made because Europe was in average ahead of time in terms of mortality rate, at least during the first COVID-19 epidemic wave.

To explain the mixture probabilities (i.e., the similarity between targets and contributors in terms of mortality dynamics), we consider 29 factors related to economy, demography, health, healthcare system and climate (see Supporting Table S9 and Soubeyrand et al., 2020a). More precisely, our objective is to identify factors negatively correlated with the response, i.e., the lower the distance between two geographic entities with respect to a given relevant factor, the higher the mixture probability. Consequently, the statistical test used is univariate test (iii) for continuous variable and the factors related to a specific pair (US state - EU country) are computed from Equation (2).

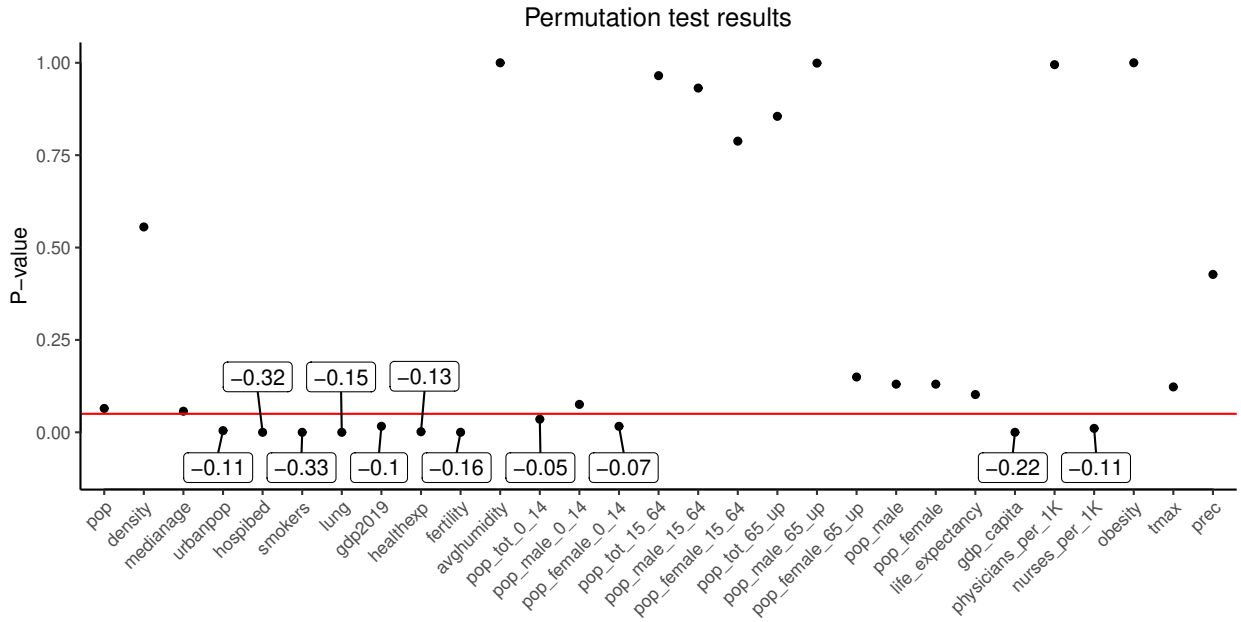

Figure 6: Comparison of the p-values (black dots) for the univariate tests (iii) associated to each factor and the significance level $\alpha = 0.05$ (red line). The Spearman correlation is given by the framed value for factors with a p-value below the threshold.

Figure 6 shows the p-values obtained for each factor and the Spearman's correlation for significant factors. We identified eleven impacting factors: hospibed, smokers, lung, healthexp, gdp_capita, fertility, urbanpop, nurses_per_1K, gdp2019, pop_female_0_14, and pop_tot_0_14. The figure shows that Spearman's correlation is negative for significant factors. This result is consistent with our objective (to identify the significant factors negatively correlated to the response). Then, we applied the multivariate analysis to the eleven impacting factors. The optimal indicator is $I_{\hat{\boldsymbol{\beta}}}(\mathbb{X}, Z) = 0.73$. The indicator value shows that a high monotonous dependency exists between the probability and

these factors.

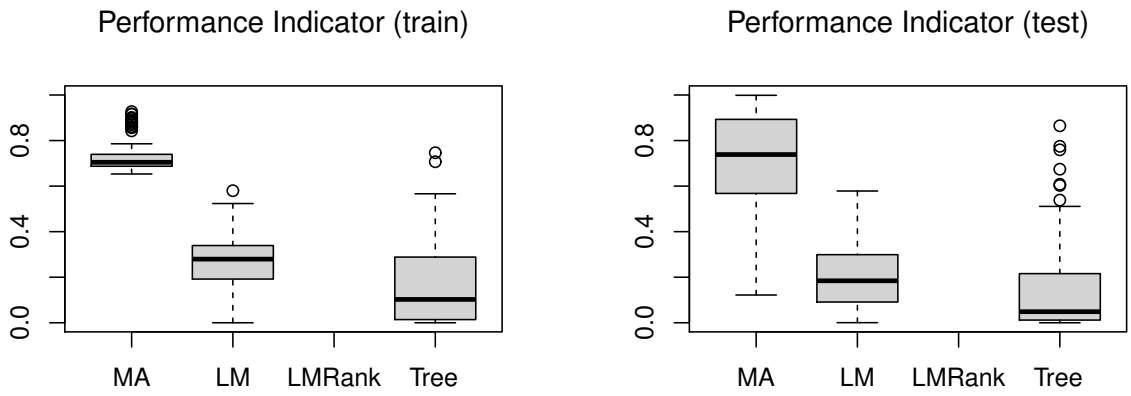

Figure 7: Boxplots of the performance indicator calculated from the train and test samples for MA, LM and Tree for the COVID-19 study.

Finally, a cross-validation step described in sub-section 2.4 is realized to ensure the robustness of the methodology. Figure 7 and S8 shows the good performance of the multivariate analysis (MA) compare to linear regression (LM) and decision tree (Tree). In addition, the two indicators remains stable along train samples and test samples. We observe that the two indicators have very similar values.

In conclusion, the factors hospibed, smokers, lung, healthexp, gdp_capita, fertility, urbanpop, nurses_per_1K, gdp2019, pop_female_0_14, and pop_tot_0_14 have a high monotonous link with the similarity measure between US states and EU countries. Our methodology gives very good results for the orderly classification of contributors when the classical prediction methods (LM, LMRank and Tree) not. The boxplot for the performance indicator of LMRank method is absent because the prediction gives only null probabilities (the calculation of the correlation is therefore impossible). The cross validation shows that the method developed in this paper can be used to reliably predict the rank order of contributor countries for each target country when correlated factors have been previously identified.

5 Discussion

In this paper we propose a methodology to deal with zeros-inflated proportion response with dependency structure. The proposed method is validated on simulations. We apply each proposed test on continuous and discrete factors. We calculate the type I and type II errors. On average, the type I error is less than or equal to the threshold. The type II error is very satisfactory for all the tests (< 0.2) but is smaller for the tests on discrete factors. Therefore, false positives are very well controlled for all tests. False negatives are very well controlled for discrete factors and sufficiently controlled for continuous factors. We then calculated the evolution of the performance indicator following the sequential and orderly addition of significant factors. The results show that the average optimal indicator is obtained when all significant factors are added. Moreover, the indicator is robust to the addition of non-significant factors. The values of the parameters in the linear combination correspond on average to the simulated parameters. In order to estimate these parameters, we used a time-consuming genetic algorithm. It would be interesting to reduce this computational cost by using for example linear algebra and analysis tools as presented in the paper of Alfons et al. (2016). Finally, the cross validation shows that our method gives slightly better results than linear regressions in this context of simulations. In conclusion, the developed methodology is validated and can be applied on the two real data sets.

The first concerns the Equine Influenza epidemic in New Market in 2003 and the second the Covid19 pandemic around the world. The results obtained in the case of equine influenza allowed us to confirm the importance of direct contact between hosts for the virus transmission. Indeed, the more horses were in regular contact (close or identical yard), the higher the probability of transmission. Concerning the factors related to sex, the interpretation is more complex. We have ruled out the possibility of a confounding effect with a factor presented in our data set. In view of the results, we are leaning towards an environmental explanation (transport, groom, jockey, ...), behavioral or linked to the immune response. Moreover, the low value of the indicator (0.21) shows that there are certainly other determining factors to explain the transmission probabilities. Validating or invalidating these hypotheses on other equine influenza datasets would be very interesting to improve the prevention of this virus. Finally the cross validation study shows the stability of the method regardless to the prediction method. In this application, the decision tree gives the best performance indicator. The decision tree and the multivariate analysis gives the same performance to identify the contributors with the higher probabilities for each target. However, the value of the two indicators are too low to give an accurate prediction of the ranks of the contributors regardless of the chosen method. The results obtained on the probabilities of similarity of the covid19 mortality curve show that certain macroscopic factors are correlated with these probabilities. We find three factors related to demography (urban population, population and women under 14 years old) and two related to GDP. We also find many factors related to the health of the population (smokers, mortality rates related to lung disease, fertility rates) but also to the means put in place by the state for health (expenditures, nurses and number of hospital beds). The high value of the performance indicator (0.73), allows us to validate this set of factors even if it remains an unexplained part of the rank correlation. Finally, the cross-validation study confirms the robustness of this study. The results of the comparison with the decision trees and the linear regression model confirm the necessity of the developed method. These results may allow each state to find several benchmark countries that are similar to it in terms of factors. Then, state could adopt, based on the results of the benchmarked countries, a similar or dissimilar strategy in the management of the epidemic. We obtained these results for probabilities calculated only on June 6, 2020. It would be interesting to study the evolution of the factors potentially impacting at different times of the epidemic.

In conclusion, the proposed method allows the reliable identification of factors correlated with a zeros-inflated proportion response with dependency structure. In addition, the cross-validation steps applied in three very different contexts show that the prediction of ranks using the performance indicator gives stable and promising results.

Supporting information

S1 Why by-block permutations?

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 $n_t n_c$ realizations of a factor X such that $x^{1,1} >$

$x^{1,2} > \ldots > x^{n_t,n_c}$ and $n_t n_c$ realizations of the response Z such that for a fixed receiver i:

$$z_1^i \leq \ldots \leq z_{n_c}^i \tag{S1}$$

$$\sum_{j=1}^{n_c} \mathbf{1}_{z_j^i > 0} = c, c \leq n_c \tag{S2}$$

$$\sum_{j=1}^{n_c} z_j^i = \frac{i}{n_t} \tag{S3}$$

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 target and contributing nodes are placed as illustrated in Figure S1. The response are the probabilities of transmission and the factor is the distance between hosts. The closer is a contributor to a target, the higher is the probability of transmission (Equation (S1)). Only a given number of hosts are potential contributors (Equation (S2)). The Equation (S3) can comes from an external contributor 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.

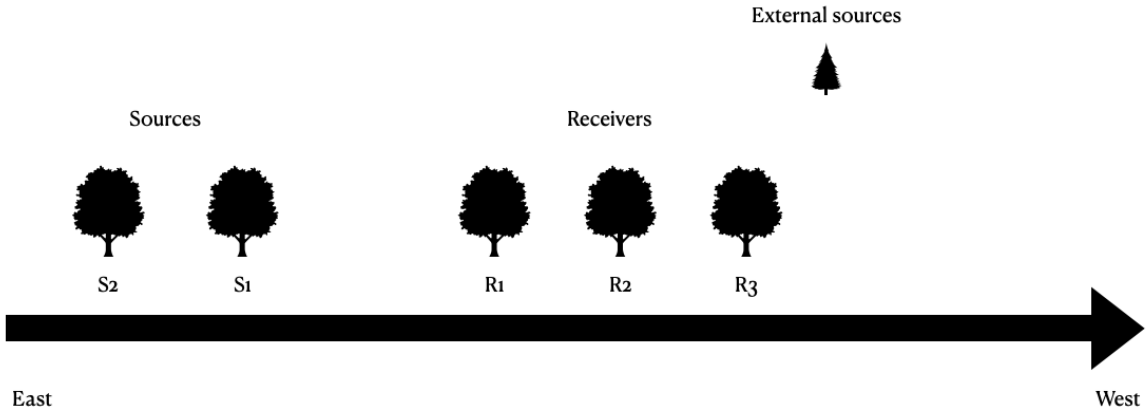

Figure S1: Schematic representation of the position of the trees.

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 $n_t = 10$, $n_c = 20$ and $c = 5$ with 1000 simulated responses. The response is computed as follows, $\forall i \in \{1, \ldots, n_t\}$:

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: $(z_1^i, \ldots, z_{n_c}^i) = \frac{i}{n_t \sum_{k=1}^{c} y_k}(0, \ldots, 0, y_c, \ldots, y_1)$

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

The estimate power of the by block-permutations tests is 1 and 0.05 without block (at $\alpha = 0.05$).

In conclusion, the by block-permutations are crucial to identify factors that are correlated to the response variable.

S2 $\Delta_{\mathbf{x},\mathbf{y}}$ calculation

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 assumed 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 $\mathbf{x}$:

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

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

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

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

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

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.

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

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

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

S3 Estimated errors of one-tailed permutation tests

Table S1: Estimated type I errors of the one-tailed (ii) permutation tests with 1000 repetitions.

	Continuous factors					Discrete factors				
m	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$	$\mathbf{x}_{16}$	$\mathbf{x}_{17}$	$\mathbf{x}_{18}$	$\mathbf{x}_{19}$	$\mathbf{x}_{20}$
0.1	0.055	0.044	0.049	0.033	0.045	0.057	0.041	0.048	0.049	0.053
0.15	0.041	0.048	0.048	0.040	0.054	0.059	0.046	0.045	0.043	0.050
0.2	0.059	0.049	0.051	0.060	0.049	0.054	0.040	0.051	0.063	0.054
0.25	0.050	0.053	0.062	0.042	0.046	0.054	0.039	0.045	0.038	0.055

Table S2: Estimated type I errors of the one-tailed (iii) permutation tests with 1000 repetitions.

	Continuous factors					Discrete factors				
m	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$	$\mathbf{x}_{16}$	$\mathbf{x}_{17}$	$\mathbf{x}_{18}$	$\mathbf{x}_{19}$	$\mathbf{x}_{20}$
0.1	0.052	0.044	0.047	0.041	0.054	0.067	0.038	0.038	0.058	0.045
0.15	0.034	0.050	0.059	0.054	0.064	0.054	0.048	0.048	0.053	0.044
0.2	0.058	0.053	0.047	0.055	0.055	0.054	0.040	0.052	0.060	0.045
0.25	0.048	0.051	0.065	0.039	0.047	0.052	0.048	0.058	0.048	0.051

Table S3: Estimated type II errors of the one-tailed (ii) permutation tests with 1000 repetitions.

	Continuous factors		Discrete factors	
m	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_{14}$	$\mathbf{x}_{15}$
0.1	0.116	0.129	0.000	0.000
0.15	0.062	0.069	0.000	0.000
0.2	0.069	0.053	0.000	0.000
0.25	0.042	0.043	0.000	0.000

Table S4: Estimated type II errors of the one-tailed (iii) permutation tests with 1000 repetitions.

	Continuous factors			Discrete factors		
m	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$
0.1	0.113	0.127	0.119	0.000	0.000	0.000
0.15	0.081	0.097	0.079	0.000	0.000	0.000
0.2	0.057	0.057	0.045	0.000	0.000	0.000
0.25	0.031	0.047	0.030	0.000	0.000	0.000

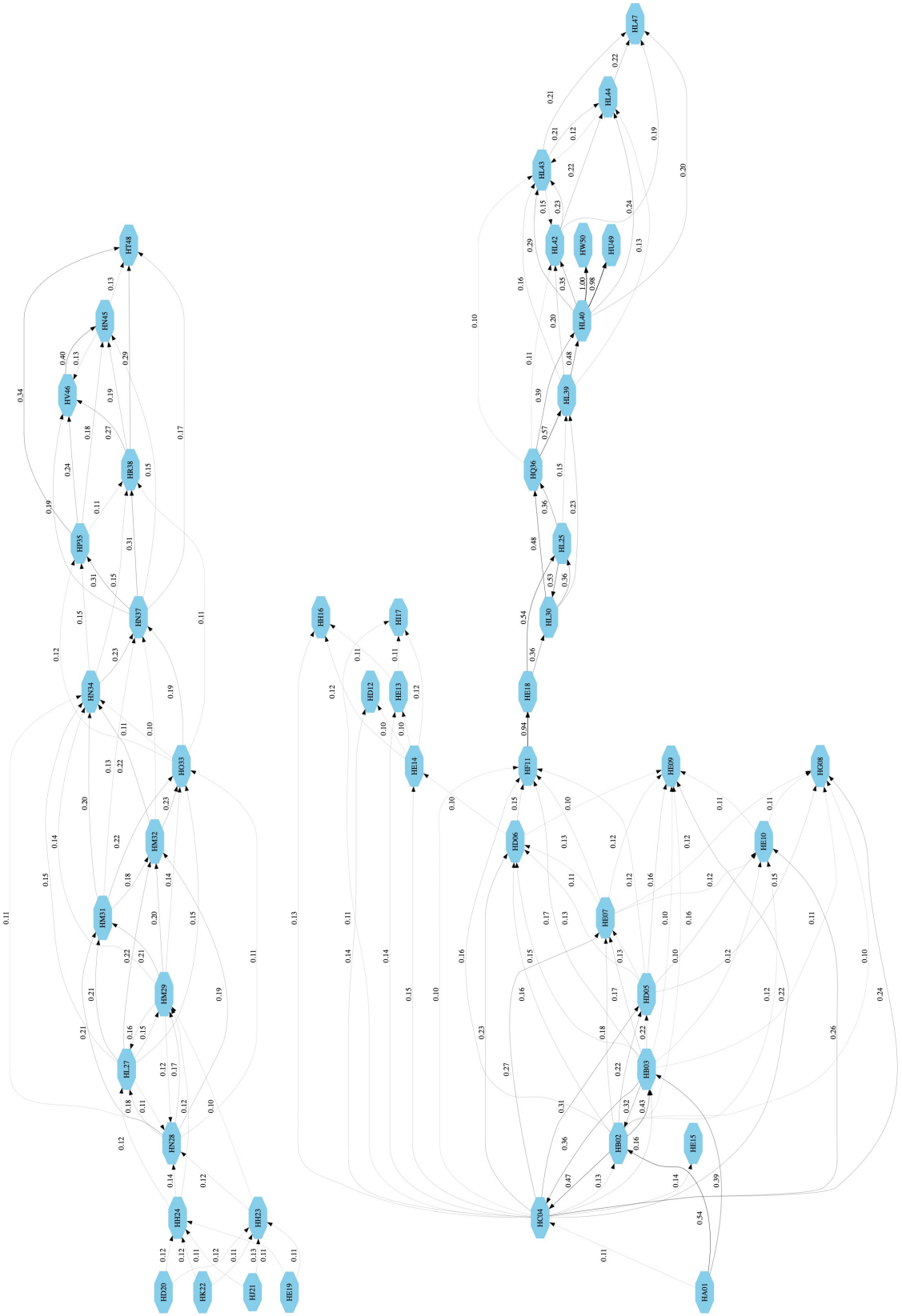

Figure S2: Transmission tree for the equine influenza outbreak inferred with BadTrIP. Blue ellipses: hosts; arrows: transmission links; Figures accompanying arrows: transmission probabilities. aïe mes yeux ;-)

Table S5: Description of equine influenza data. Left: Counts of hosts with non-missing information for each observed variable. Right: counts of potential source-receptor pairs without non-missing information for pairwise factors.

	#
All	48
Yard	48
Age	27
Sex	26 (9 females, 17 males)

	#pairs
All	2256
Same_Yard	2256
Dist_Yard	2256
Diff_Age	702
Same_Sex	650
Trans_Sex	650

S4 Exploration of the significance of factors related to sex

Here, we investigate eventual confounding effects related to the significant effects of "Same_Sex" and "Trans_Sex" factors on the transmission probability. Even if our permutation tests do not require balanced classes, we firstly explore whether the trend for higher probabilities of F→F transmissions coincides with an excess of female horses. Actually, the number of females is about the half of the number of males (Figure S3, left). Therefore, under complete (uniform) randomness, we would expect about two times more M→F transmissions than F→F transmissions (and two times more M→M than F→M). When we only consider the occurrences of "Trans_Sex" corresponding to positive probabilities (without accounting for null probabilities), we clearly see the excess of F→F and F→M transmissions compared to their expected values under complete randomness (Figure S3, right). To complete this observation, transmission probabilities were inferred to be positive for only 19% of all the possible M→F pairs (0.7 times less than expected under complete randomness), 53% for F→F (1.8 times more than expected under complete randomness); see Table S6. Hence, gender distribution is not likely to be involved in the significant effect of the factors related to sex.

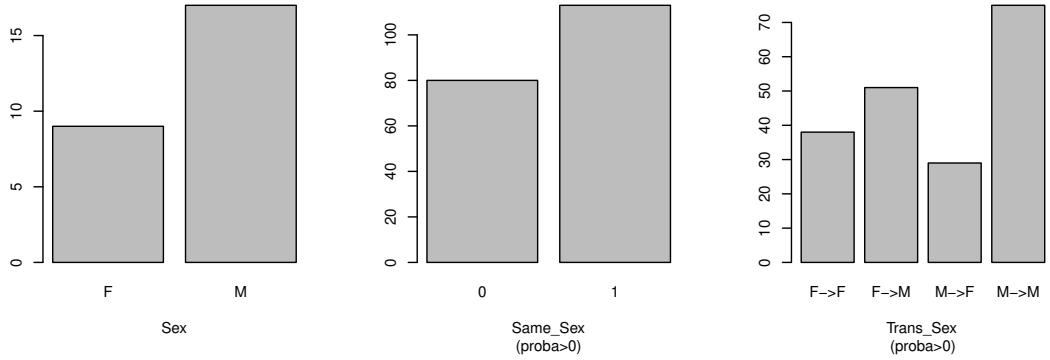

Figure S3: Left: gender distribution (female and male) in the equine influenza study. Center and right: distributions of the variables "Same_Sex" and "Trans_Sex" corresponding to pairs associated with positive transmission probabilities.

Table S6: Statistics about "Trans_Sex" modalities for the equine influenza study. Line 1: Transmission probabilities inferred to be positive for all the possible source-receptor pairs with respect to each modality of the "Trans_Sex" variable. Lines 2 and 3: Risk ratio and odds ratio, respectively, for each "Trans_Sex" modality between the observed situation and the hypothetical case of completely random transmissions.

Statistic	F→F	F→M	M→F	M→M
% of positive proba.	53	33	19	28
Risk ratio	1.8	1.2	0.7	1.0
Odds ratio	2.9	1.3	0.4	0.9

Secondly, we explore the eventual existence of confounding factors among those we have considered. As shown by Figure S4, "Same_Sex" and "Trans_Sex" are not correlated with "Same_Yard" and "Dist_Yard", and "Trans_Sex" is only slightly correlated with "Diff_Age". The absence of link between the two yard variables and the two gender variables is confirmed by Figure S5 and Table S7. Hence, there seems to be no confounding factors in the data set, and the significance of gender-related factors has to be explained by external processes (e.g., the indirect contacts between hosts via groom, jockey or transport, the behavior of horses in herds, or different immune responses depending on the sex).

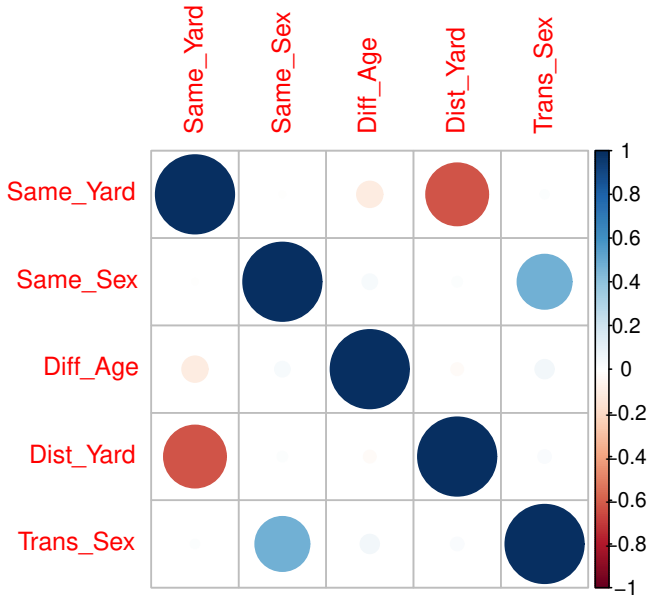

Figure S4: Matrix of Pearson's correlations between factors in the equine influenza study.

Table S7: Result of the independence chi-squared test applied to qualitative factors considered in the equine influenza study.

Factors	χ^2-test statistic	p-value
Same_Yard : Same_Sex	0.04	0.85
Same_Yard : Trans_Sex	1.76	0.62
Same_Yard : Diff_Age	4.3	0.12
Same_Sex : Trans_Sex	650	$< 2.2e^{-16}$

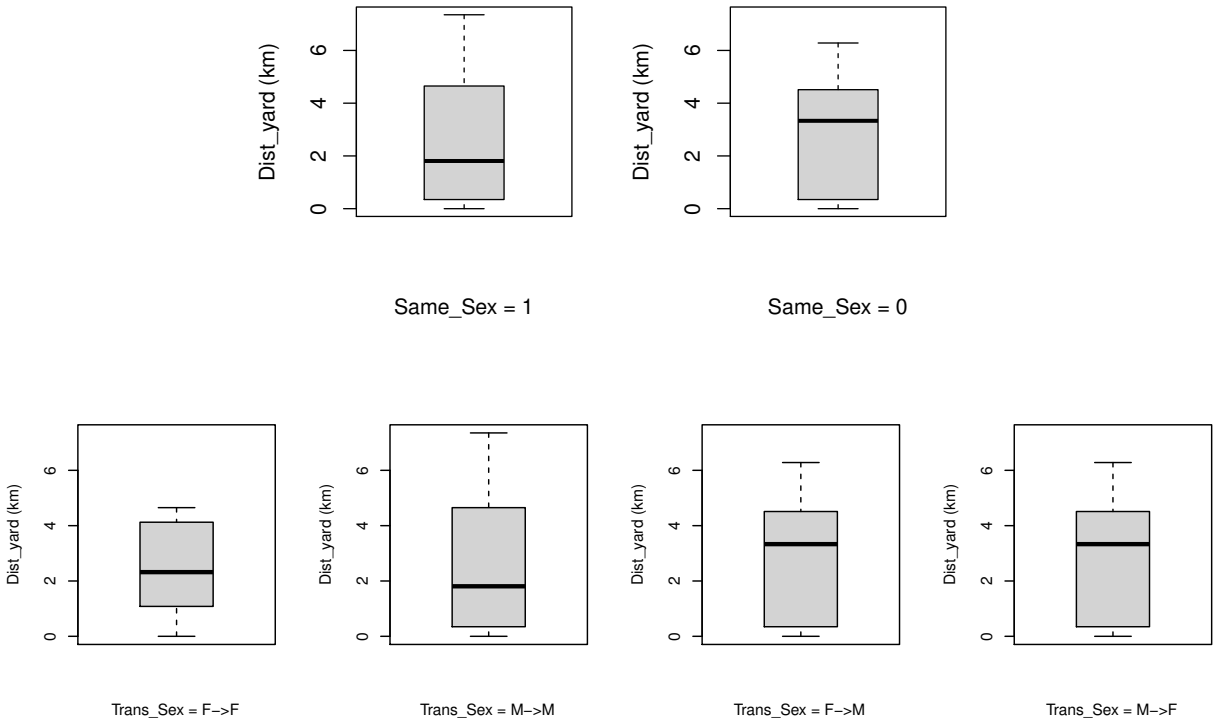

Figure S5: Distribution of the variable "Dist_Yard" by modality of the variable "Same_Sex" and "Trans_Sex" in the equine influenza study.

S5 Cross validation method to compare multivariate analysis (MA), linear regression (LM) and regression tree (Tree)

The objective of the cross validation step is to compare our methodology to the linear regression and the decision tree. The least square error is minimized on each train set for the additive linear regression. The L2 norm of least squares estimation is replaced by a pseudo-norm which is a function of the ranks of the residuals for the additive linear regression based on rank. 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)).

In this cross validation, we use two indicators to compare our methodology. The first indicator is the performance one. The performance indicator is equal to $I(\hat{\mathbf{z}}, \mathbf{z})$ where $\hat{\mathbf{z}}$ is equal to $M_{\mathbb{X}}\boldsymbol{\beta}$ for multivariate analysis (MA) and is the prediction of the response computed for linear regression (LM) and decision tree (Tree). The second indicator is related to the rank of the target-contributor couples that has a strictly positive probability. For a fixed target i, we order the contributors such that $\left(\hat{R}^{\ell=1}_{\mathbf{z}_i^j} < \hat{R}^{\ell=2}_{\mathbf{z}_i^j} < \ldots < \hat{R}^{\ell=n_c}_{\mathbf{z}_i^j}\right)$ and we compute:

$$CR_i = \frac{1}{\#E_i} \sum_{\ell=n_c-\#E_i}^{n_c} \mathbf{1}(\ell \in E_i)$$

where $E_i = \{j | z_j^i > 0\}$ and $\hat{R}_{\mathbf{z}_i^j}$ is the predicted rank of the couple (i, j). For example, let 10 contributors $(1, \ldots, 10)$ and the fixed target i. There is only the first three couples with a strictly positive probability. The probabilities are given by the first row of the Table S8. The second row gives the true ranks and the three last rows some possible predicted ranks. The indicator CR_i is shown on the last column.

Table S8: An illustrative example for the computation of the indicator CR_i.

	1	2	3	4	5	6	7	8	9	10	CR_i
z_j^i	0.3	0.2	0.3	0	0	0	0	0	0	0	
$\hat{R}_{z_j^i}$	9	8	10	4	4	4	4	4	4	4	1
$\hat{R}_{z_j^i}$	9	10	8	3	7	3	3	6	3	3	1
$\hat{R}_{z_j^i}$	7	10	8	3	9	3	3	6	3	3	2/3
$\hat{R}_{z_j^i}$	6	4	7	4	9	10	4	8	4	4	0

This indicator quantifies the quality of identification of the contributors with the highest ranks. The order of these within the group does not matter and the order of others outside the group does not matter.

Finally the Contributor Ranking (CR) indicator is:

$$CR = \frac{1}{n_t} \sum_{i=1}^{n_t} CR_i$$

In conclusion, the performance indicator gives an idea of the method's ability to order probabilities globally. The second indicator (CR) focuses on the positive probabilities per target. It is very useful in the context of our applications. Indeed, the higher the indicator, the more the method will be able to give with certainty the potential contributors for each target.

The following figures show the CR criterion calculated for the four models on the training sets and on the test sets for the simulations and two application cases (equine influenza and covid-19).

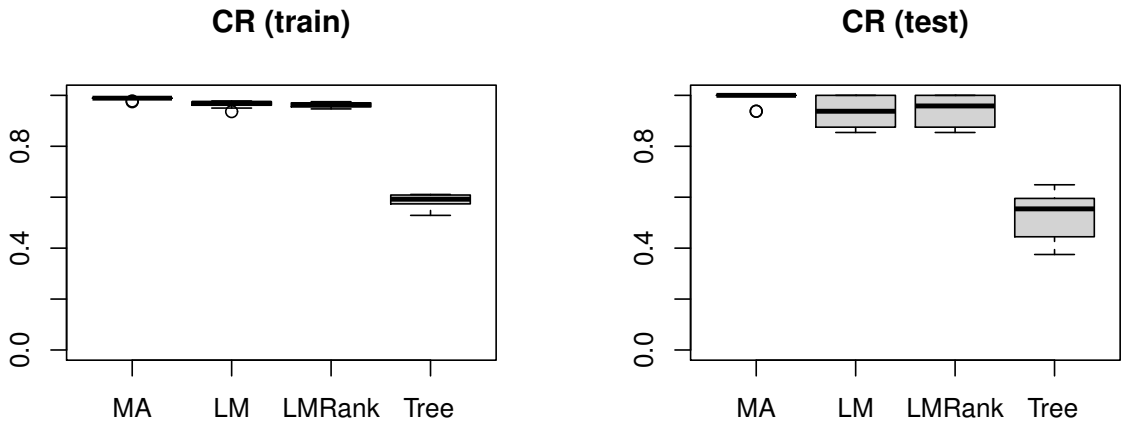

Figure S6: Boxplots of the contributor ranking indicator calculated from the train and test samples for MA, LM and Tree in the simulations.

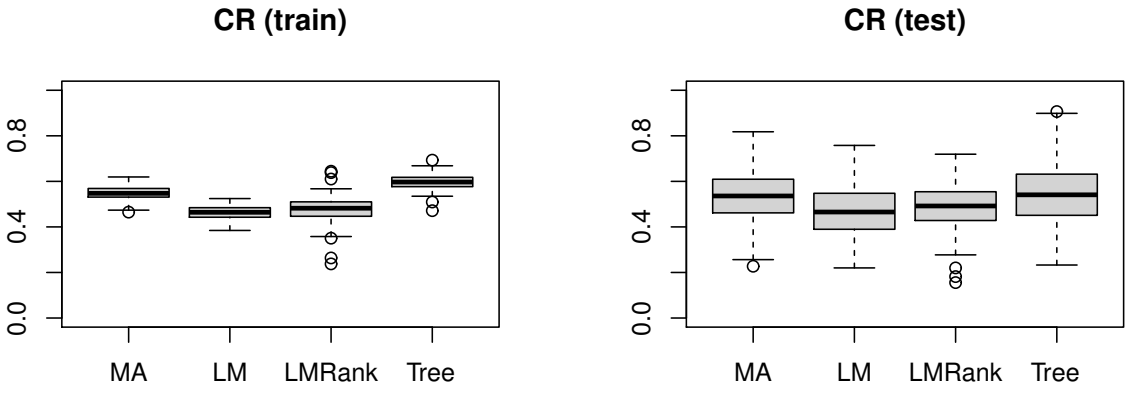

Figure S7: Boxplots of the contributor ranking indicator calculated from the train and test samples for MA, LM and Tree in the equine influenza study.

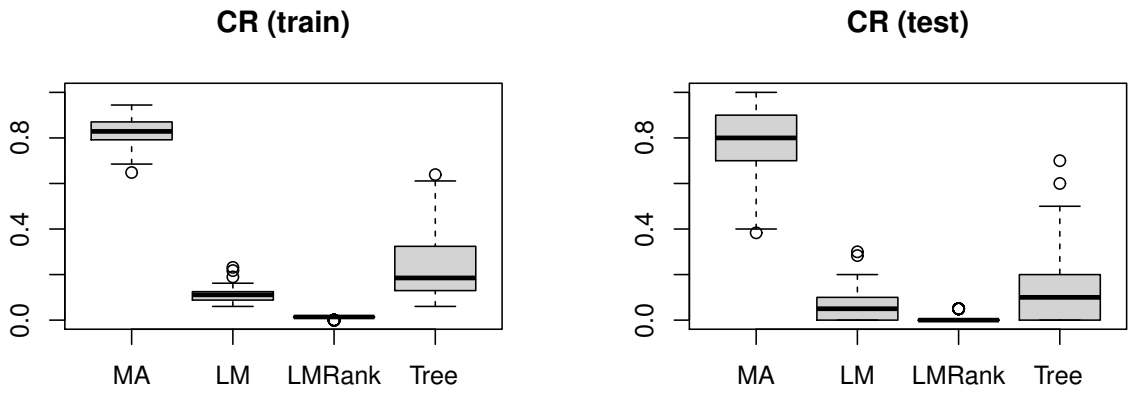

Figure S8: Boxplots of the contributor ranking indicator calculated from the train and test samples for MA, LM and Tree for the COVID-19 study.

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

Table S9: 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	%

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. (2020). Reconstruction des transmissions d'un virus au cours d'une épidémie par apprentissage statistique sur données génomiques. PhD thesis, Aix-Marseille Université, Marseille.

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.

Alfons, A., Croux, C., and Filzmoser, P. (2016). Robust maximum association between data sets: The r package ccapp. 45:71–79.

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

De Maio, N., Worby, C. J., Wilson, D. J., and Stoesser, N. (2018). Bayesian reconstruction of transmission within outbreaks using genomic variants. PLoS computational biology, 14(4):e1006117.

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.

Dunn, O. J. (1964). Multiple comparisons using rank sums. Technometrics, 6(3):241–252.

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.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. Biometrika, 75(4):800–802.

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.

Kloke, J. D. and McKean, J. W. (2012). Rfit: Rank-based estimation for linear models. The R journal, 4(2):57–64.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American statistical Association, 47(260):583–621.

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.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(3):507–554.

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 Roques, L. (2020b). COVID-19 mortality dynamics: The future modelled as a (mixture of) past (s). Plos one, 15(9):e0238410.

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.

Tsagris, M. and Stewart, C. (2018). A Dirichlet regression model for compositional data with zeros. Lobachevskii Journal of Mathematics, 39(3):398–412.

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