Identifying potential significant factors impacting zero-inflated proportions data

Melina Ribaud, Edith Gabriel, Joseph Hughes, Samuel Soubeyrand

Identifying potential significant factors impacting zero-inflated proportion data

Mélina Ribaud[1,*], Edith Gabriel[1], Joseph Hughes[2], and Samuel Soubeyrand[1,*]

[1]INRAE, BioSP, 84914 Avignon, France
[2]MRC-University of Glasgow, Centre for Virus Research, Glasgow, Scotland, United Kingdom
[*]Corresponding authors: melina.ribaud@gmail.com and samuel.soubeyrand@inrae.fr

April 3, 2023

Abstract

Classical supervised methods like linear regression and decision trees are not completely adapted for identifying impacting factors on a response variable corresponding to zero-inflated proportion data (ZIPD) that are dependent, continuous and bounded. In this article we propose a within-block permutation-based methodology to identify factors (discrete or continuous) that are significantly correlated with ZIPD, we propose a performance indicator quantifying the percentage of correlation explained by the subset of significant factors, and we show how to predict the ranks of the response variables conditionally on the observation of these factors. The methodology is illustrated on simulated data and on two real data sets dealing with epidemiology. In the first data set, ZIPD correspond to probabilities of transmission of Influenza between horses. In the second data set, ZIPD correspond to probabilities that geographic entities (e.g., states and countries) have the same COVID-19 mortality dynamics.

Keywords: COVID-19; Equine Influenza; Performance Indicator; Permutation Test; Ranking; Spearman's correlation.

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 may 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). The network vision of data is illustrated by two schematic configurations shown in Figure 1, which includes notations that are introduced in Section 2, and our general objective in this paper is to unravel which factor(s) characterizing the pairs of nodes *explain* the network edges.

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 a health crisis. Indeed, assessing the influence of social, biological and environmental factors in the spread of epidemics contributes to identifying mechanisms for controlling the disease dynamics. The spread of epidemics can be understood by quantifying 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[1], and similarity measures of disease dynamics in several geographic entities[2]. 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 because dependencies and overdispersion may lead to biased results if ignored[3][4]. Here, we focus on epidemiological links defined as proportions and we aim to provide a statistical methodology for explaining these 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[5], the linear regression model[6] predicts response values and identifies influencing factors. When the response variable follows other frequently used distributions (e.g., binomial or Poisson), the generalized linear models[7] (GLM) can be considered. When one prefers to avoid making a distributional assumption for the response variable, non-parametric predictive models[6] may be a solution. However, non-parametric models do not standardly provide direct testing procedure to identify impacting factors.

In the exploration of the link between factors and a response variable, we are particularly interested in zero-inflated response variables (i.e., random variables with a positive mass at zero, or, in other words, variables presenting an excess of zeros). In classification problems (i.e., problems with categorical response variables), excess of any class can be handled by balancing classes using resampling methods[8]. In regression problems (i.e., problems with quantitative response variables), zero-inflation is typically handled by defining a model as a mixture of two processes: the first process generating only zeros, the second process being governed by a usual distribution such as the zero-inflated Poisson, zero-inflated binomial and zero-inflated beta distributions[9]. For such zero-inflated models, generally assuming independence between observations, the influencing factors can be identified with statistical tests[10].

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[11], that have been classified with respect to the nature of the response[12] (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, an adaptation of the zero-inflated Dirichlet regression (ZIDR) model was proposed for microbiome compositional data[13].

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 two levels. The GLM (including zero-inflated data) and the ZIDR can treat continuous factors as well as discrete factors with only two levels, even if this restriction is minor since a factor with multiple levels can be treated as several factors with two levels each.

Here 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. 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 mutually 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 precisely a distribution-free) approach based on permutation tests aiming (*i*) to identify factors (discrete or continuous; characterizing the target, the contributor or the target-contributor pair) that are significant, (*ii*) to quantify via a performance indicator the percentage of correlation explained by the subset of significant factors, and (*iii*) to predict the ranks of proportions from the significant factors. To take into account the dependence of proportion data linking the contributing nodes to any target node, we define a test grounded on the principle of within-block permutations[14,15,16]. In this test, permutations are constrained by the dependence structure of data by shuffling proportion data only within each 'contributors–target' block (for any target). In addition, the test deals with both factor types, continuous and discrete, by adapting the test statistic, which 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. The zero-inflation is not explicitly handled in the test itself (permutation procedures are nevertheless considered as versatile for variables clumping at zero because they do not require distributional assumptions[17]), but it is accounted for in the performance indicator whose range is robust to the proportion of zeros (and more generally to the proportion of ties). Supporting Information, Figure S1, gives a general picture about the workflow of the methodology briefly described above.

In what follows, Section 2 sets the framework and notations and presents the procedure based on permutation tests to identify the factors correlated to the response as well as the performance indicator. The method is then applied to simulations (Section 3) and to real data dealing with Equine Influenza and COVID-19 epidemics (Sections 4 and 5).

Table 1: Comparison of models in their ability to match the constraints considered in this article.

Methods	Response			Factor		Dependency
	Distribution free	$[0,1]$	Zero inflated	Tests: Discrete	Tests: Continuous	
Linear regression[6]				✓[a]	✓	
Beta regression[9]		✓	✓	✓	✓	
Dirichlet regression[18]		✓	✓	✓	✓	✓
Decision tree[19]	✓	✓				✓

[a] ANOVA and ANCOVA

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 target $i \in \{1, \dots, n_t\}$ and contributor $j \in \{1, \dots, n_c\}$, as illustrated in Figure 1. The higher the value, the stronger the link and the weaker the other links. We assume that Z_j^i is continuous, $Z_j^i \in [0, 1]$ and, for any target node, the sum over all contributors cannot exceed 1, i.e.:

$$\sum_{j=1}^{n_c} Z_j^i \leq 1, \ \forall i \in \{1, \dots, n_t\}. \tag{1}$$

In addition, in typical applications we will consider that the distribution of Z_j^i is zero-inflated.

The factors characterize any target-contributor pair (i.e., the two nodes and their interaction), as illustrated in Figure 1. 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 are often grounded on information separately collected from the target node i and the contributing node j. In this case, $x_k^{(i,j)}$ is defined from any application taking as arguments the values of a variable observed from the contributor and target nodes. Intuitive examples include the geographical distance measured from the spatial coordinates of the contributor and the target, the absolute difference between the ages of the contributor and the target, etc. Further examples provided in the application sections lead to discrete, continuous and even categorial factors.

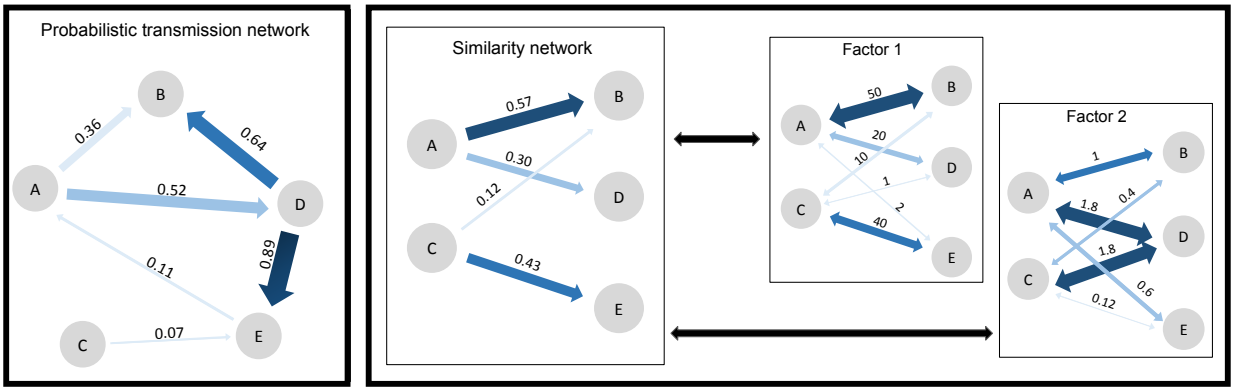


Figure 1: Typical configurations of epidemiological networks of interest and links to factors. Left: Probabilistic transmission network; Arrows indicate probable transmissions (with associated probabilities given by numbers above the arrows); Host C is a contributor (i.e., probable source of infection), host B is a target (i.e., an infection receptor) and hosts A, D and E are both contributors and targets (thus, $n_t = n_c = 4$); E.g., A infected B with probability $Z_{\mathrm{A}}^{\mathrm{B}} = 0.36$. Right: Similarity network; Arrows indicate probabilities that targets (namely host units B, D and E; $n_t = 3$) have the same features as (e.g., 'follow the mortality dynamics of') contributors (namely hosts units A and C; $n_c = 2$); E.g., B follows the mortality dynamics of A with probability $Z_{\mathrm{A}}^{\mathrm{B}} = 0.57$. Notes: The objective of our approach is to assess the link between an epidemiological network and multiple factors that characterize pairs of host units, as illustrated on the right-hand side where Factor 1 seems to better reflect the similarity network than Factor 2 does (the numbers above the arrows for the panels representing Factors 1 and 2 are the values taken by these undirected factors); In the main text, probabilities Z_j^i are indexed by host indices instead of host names; The sum of inward probabilities for a given target is lower than 1 if some contributors are unobserved.

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

The specific characteristics of our response variable make impossible the use of classical correlation tests such as Spearman's test[20], because the response has numerous ties (zeros). Solutions to treat ties in ranking problems were proposed[21], but for cases with numerous ties, 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. Within-block permutation tests, grounded on the assumption of exchangeability of data within blocks[14,15,16], appear to be a possible alternative to take into account these constraints.

Let $\mathbf{x}_k \in \mathbb{R}^{n_t n_c}$, $k = 1, \ldots, d$, 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[22] for block-permutation to test the correlation between the response and the factor. We denote by T_k the test statistic, which is dependent on the type factor, and by $\lambda_k(\mathbf{z})$ the p-value. The CMC algorithm for block-permutation test adapted to data that we consider in this article consists of the following steps:

1. Compute the statistic T_k on the original data set $(\mathbf{x}_k, \mathbf{z})$;
2. Do B independent repetitions of what follows: randomly permute the response by block of type target–contributors, set the new response vector denoted $\mathbf{z}^b$, $b = 1, \ldots, B$, and compute the statistic T_k^b on the permuted data set $(\mathbf{x}_k, \mathbf{z}^b)$;
3. Estimate the p-value by $\widehat{\lambda}_k(\mathbf{z}) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{\{T_k^b \geq T_k\}}$.

Remarks: The T_k statistic must be positive to calculate the p-value using the CMC algorithm. The permutation by block of type target–contributors is carried out by permuting the components of the vector $(z_1^i, \ldots, z_{n_c}^i)$ for each target node i (the permuted response $\mathbf{z}^b$ hence satisfies mutual dependencies between its components summarized in the system of n_t constraints given by Equation (1), and keeps eventual heterogeneity in the distribution characteristics of $(z_1^i, \ldots, z_{n_c}^i)$ between targets). Block permutations are required to minimize the second species risk of the test (see Supporting Information, Appendix A).

For a continuous factor $\mathbf{x}$, by omitting the subscript k, the statistic T is the non-parametric Spearman's correlation[23] between $\mathbf{x}$ and $\mathbf{z}$, say $r_s(\mathbf{x}, \mathbf{z})$, i.e. it is defined as the Pearson correlation between the rank variables $r_s(\mathbf{x}, \mathbf{z}) = \rho(R_\mathbf{x}, R_\mathbf{z})$, where ρ is the Pearson correlation, $R_\mathbf{x}$ (resp. $R_\mathbf{z}$) is the random vector that gives the ranks of the elements of $\mathbf{x}$ (resp. $\mathbf{z}$). 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 the test statistic is $T = r_s^2(\mathbf{x}, \mathbf{z})$;

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

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

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[24]:

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

where $R_{z_j^i}$ denote the rank of the element (i,j) of $\mathbf{z}$, $\bar{R}_{\mathbf{z}} = \frac{1}{n_t n_c}\sum_{i=1}^{n_t}\sum_{j=1}^{n_c} R_{z_j^i}$, $n_q = \sum_{i=1}^{n_t}\sum_{j=1}^{n_c} \mathbf{1}_q(x^{(i,j)})$, $\bar{R}_{\mathbf{z}_{.q}} = \frac{1}{n_q}\sum_{i=1}^{n_t}\sum_{j=1}^{n_c} R_{z_j^i}\mathbf{1}_q(x^{(i,j)})$ 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 of the factor being considered is less than 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 $T = (\bar{R}_{z_{.q}} - \bar{R}_{z_{.\tilde{q}}})^2$;

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

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

Notes: Here, the statistics are the differences in mean ranks[25]. 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 and sharper improved Bonferroni correction called Benjamini-Hochberg correction[26,27]. As an illustration, we provide both the Bonferroni and the Benjamini-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

To take into account the multivariate aspect of the correlation, we develop a performance indicator that simultaneously accounts for all discrete and continuous factors previously identified. The indicator can be viewed as a surrogate for the coefficient of determination used in linear regression, representing the monotonous relationship between a single linear combination of all factors and the response. It varies in $[0,1]$; the closer to 1, the stronger the correlation between the ranks of the *best* combination of the set of factors $\mathbb{X}$ (which corresponds to the term $M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$ defined below) and the ranks of the response variable $\mathbf{z}$. To ensure that the performance indicator can effectively reach the maximum value 1, it is defined as the ratio between the Spearman correlation and its actual upper bound. The upper bound is computed analytically by following the reasoning proposed by Kendall[21] for treating ties in ranking problems. The reasoning adapted to our problem consists in identifying the situation where the Spearman correlation is maximum for given $\mathbf{z}$ and $M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$ ignoring the actual pairing between these sets of variables. This situation occurs when the ranks of

the (i,j)-th components of $M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$ and $\mathbf{z}$ are equal for any (i,j) such that $z_j^i \neq 0$. Under this assumption, the computation of the Spearman correlation simplifies and one obtains an explicit expression for the upper bound. In what follows, we derive the expression of the performance indicator; details on the computation of the upper bound of the Spearman correlation are provided in Supporting Information, Appendix B.

The performance indicator is built from the following function of β:

$$I_{\boldsymbol{\beta}}(\mathbb{X},\mathbf{z}) = r_s^2(M_{\mathbb{X}}\boldsymbol{\beta},\mathbf{z})(1+\Delta_{M_{\mathbb{X}}\boldsymbol{\beta},\mathbf{z}}), \tag{3}$$

where $(1+\Delta_{\mathbf{x},\mathbf{y}})^{-1}$ is the upper bound of $r_s^2(\mathbf{x},\mathbf{z})$ and $\Delta_{\mathbf{x},\mathbf{y}} = \dfrac{\sum_{i\in I_0}(R_{x_i}^2 - R_{y_i}^2)}{(n-1)\hat{\sigma}_{R_{\mathbf{y}}}^2}$ for all $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n$, $\hat{\sigma}_{R_{\mathbf{y}}}^2$ being the variance of $R_{\mathbf{y}}$ and $I_0 = \{i | y_i = 0\}$, and where the elements of the design matrix $M_{\mathbb{X}} \in \mathbb{R}^{n_t n_c \times d'}$ are defined by:

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

$\ell = (i-1)n_c + j$ and k are the indices of the rows and the columns of the design matrix, respectively, $\min\{\mathbf{x}_k\}$ (resp. $\max\{\mathbf{x}_k\}$) is the minimum (resp. maximum) element of the vector $\mathbf{x}_k$, $d' = \sum_{k=1}^{d}(Q_k - 1)$, with $Q_k = 2$ 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.

We then have to estimate the set of parameters $\boldsymbol{\beta}$ which maximizes the Spearman correlation $r_s^2(M_{\mathbb{X}}\boldsymbol{\beta},\mathbf{z})$ (to obtain the *best* combination of the set of factors $\mathbb{X}$ with the form $M_{\mathbb{X}}\boldsymbol{\beta}$, as evoked above). 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 a genetic algorithm implemented in the `R` package `rgenoud`[28] (the `genoud` function in this package combines an evolutionary search algorithm with a derivative-based Newton or quasi-Newton method to solve optimization problems). Hence, we estimate $\boldsymbol{\beta}$ as follows: $\hat{\boldsymbol{\beta}} = \arg\max_{\beta\in\mathbb{R}^{d''}} r_s^2(M_{\mathbb{X}}\boldsymbol{\beta},\mathbf{z})$. The solution of this maximization 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 $I_{\boldsymbol{\beta}_0}(\mathbb{X},\mathbf{z}) = I_{a\boldsymbol{\beta}_0}(\mathbb{X},\mathbf{z})$, $\forall a \neq 0$. In practice, since the maximizer can only be identified up to a scale factor, each component of $\boldsymbol{\beta}$ is constrained within the interval $[-10,10]$, and the genetic algorithm is stopped if the value of the objective function $r_s^2(M_{\mathbb{X}}\boldsymbol{\beta},\mathbf{z})$ has not increased in the last 50 iterations (with a tolerance level equal to 0.001) or if the maximum number of iterations (set to 200) has been reached. This is specified in the `genoud` function by setting the following options: `Domains = matrix(c(-10,10),byrow=TRUE,nrow=`d''`,ncol=2)`,

`max.generations = 200`, `hard.generation.limit = TRUE`, `wait.generations = 50` and `solution.tolerance = 0.001`. Finally, we calculate the performance indicator by pluging-in $\hat{\boldsymbol{\beta}}$:

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

2.4 Relative importance of factors

The optimal parameter vector $\hat{\boldsymbol{\beta}}$ must not be directly used for assessing the effect sizes of factors since it is not unique as explained in Section 2.3 (if $\boldsymbol{\beta}_0$ is a solution of the maximization, $a\boldsymbol{\beta}_0$ is also a solution for all real value $a \neq 0$). Nevertheless, the components of $\hat{\boldsymbol{\beta}}$ can be used to compare the relative importance of factors in explaining, through $M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$, the links between targets and contributors summarized by $\mathbf{z}$. Thus, we define the relative importance of factor $k \in \{1, \ldots, d\}$ (or the level q of factor k for discrete factors) with respect to the *average* factor:

$$\tilde{e}_{k,q} = \frac{e_{k,q}}{\frac{1}{d'}\sum_{k'=1}^{d'}\sum_{q=1}^{Q_k-1} e_{k',q}}, \tag{4}$$

where $e_{k,q} = |\hat{\boldsymbol{\beta}}^{(k)}|$ and $q = 1$ if factor k is continuous ($\hat{\boldsymbol{\beta}}^{(k)}$ being the component of $\hat{\boldsymbol{\beta}}$ corresponding to the continuous factor k) and $e_{k,q} = |\hat{\boldsymbol{\beta}}^{(k,q)}|$ if factor k is discrete ($\hat{\boldsymbol{\beta}}^{(k,q)}$ being the component of $\hat{\boldsymbol{\beta}}$ corresponding to the q-th level of the discrete factor k, $q \in \{1, \ldots, Q_k - 1\}$).

Remarks: (a) $\tilde{e}_k$ is unchanged if one substitutes $a\hat{\boldsymbol{\beta}}$ for $\hat{\boldsymbol{\beta}}$ ($a \neq 0$). (b) The operator $M_{\mathbb{X}}$ in the regression $M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$ homogenizes the amplitudes of variation of the factors and, therefore, of the coefficients in $\hat{\boldsymbol{\beta}}$ which are hence comparable. (c) The relative importance of factors measured by Equation (4) has to be understood in terms of factor contributions to the regression $M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$ explaining $\mathbf{z}$. Hence, these contributions might be subject to nonlinear effects of factors or correlation between factors. One may alternatively compute a marginal indicator of the importance of each factor by calculating for example the performance indicator $I_{\hat{\boldsymbol{\beta}}}(\mathbb{X}, \mathbf{z})$ where $\mathbb{X}$ is reduced to the factor of interest. Such a proposal however requires additional computation of the optima $\hat{\boldsymbol{\beta}}$ for every factor considered individually.

2.5 Rank prediction

If one ignores the value of $\mathbf{z}$, ranks of contributors can be predicted for any target by the ranks $\hat{R}_{\mathbf{z}}$ of $M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$. In other words, the first contributor to a given target is predicted to be the contributor with the largest component of the sub-vector of $M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$ restricted to the target under focus, the second contributor corresponds the second largest component, and so on. In the applications, we compare this ranking with rank predictions obtained from linear regression and decision tree (for which ranks are computed directly from the predictions

of proportions provided by these models) and with the rank-based estimation for linear models[29], which uses a distance based on a dispersion function[30] instead of the Euclidean distance. Note that the four rank-prediction approaches are implemented with the same set of factors identified by the tests presented in Section 2.2.

We use cross-validation to compare the robustness and the quality of the performance indicator and the ranking obtained from our multitest-based multivariate analysis (MMA), the linear regression model (LM), the rank-based linear regression model (LMRank) and the decision tree (Tree). Target hosts are randomly divided into a train sample (80% of targets) and a test sample (20% of targets). Tests for factor identification are applied to the global sample (union of train and test samples), while indicators are computed separately for each sub-sample. This procedure is independently repeated 100 times.

We consider two indicators: the performance indicator defined in the previous subsection and the contributor ranking indicator (CR). The CR indicator, defined in Supporting Information, Appendix C, is the average over the targets of the proportion of the N_i contributors with positive transmission probabilities for target i that are ranked among the top N_i contributors by the predictor under consideration (MMA, LM, LMRank or Tree).

3 Simulation study

We carry out a simulation study to investigate the performance of the proposed method. All `R` codes to implement the methods have been incorporated into the package `ZIprop`, freely available on `R` CRAN (`https://cran.r-project.org/package=ZIprop`) and GitLab (`https://gitlab.paca.inrae.fr/meribaud/ziprop`).

3.1 Simulated data

We simulate data under the constraints described in Section 2.1. The algorithm applied to simulate the factors and the response is described below:

1. Set values for $n_c > 1$, $n_t > 2$, $m \in [1/n_c, 1]$ (the proportion of non-zero values for the responses z_j^i), $d > 1$, $\boldsymbol{\beta} \in \mathbb{R}^{d'}$ and $\eta > 0$.

2. 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 $n_t n_c$-tuple $\mathbf{x}_k$, $k \in \{1, \ldots, d\}$, are independently drawn from the continuous (resp. discrete) uniform distribution $U(0, 1)$ (resp. $U(\{0, 1\})$) if $\mathbf{x}_k$ is a continuous (resp. discrete) factor.

3. Generate the un-constrained response vector $\tilde{\mathbf{z}} \in \mathbb{R}^{n_t n_c}$ from the Gaussian distribution $\mathrm{N}(\mathbb{X}\boldsymbol{\beta}, \eta^2\mathbf{I})$ with mean vector $\mathbb{X}\boldsymbol{\beta}$ and diagonal variance matrix $\eta^2\mathbf{I}$ where diagonal

elements are equal to η^2; subtract the minimum value of $\tilde{\mathbf{z}}$ to each element of $\tilde{\mathbf{z}}$ to get only non-negative elements: $\tilde{\mathbf{z}} \leftarrow \tilde{\mathbf{z}} - \min \tilde{\mathbf{z}}$; set to zero the $n_0 = \lceil (1-m) n_c n_t \rceil$ lowest elements in $\tilde{\mathbf{z}}$ excluding its maximal elements for each target $i \in \{1, \ldots, n_t\}$, i.e., the n_0 lowest elements in $\tilde{\mathbf{z}} - \{\tilde{z}_j^i : \tilde{z}_j^i = \max_{j'}\{\tilde{z}_{j'}^i\}\}$ ($\lceil \cdot \rceil$ is the ceiling function).

4. Compute the elements z_j^i of the response vector $\mathbf{z}$ by scaling $\tilde{z}_j^i$ for each target $i \in \{1, \ldots, n_t\}$: $z_j^i = \frac{\tilde{z}_j^i}{\sum_{j'=1}^{n_c} \tilde{z}_{j'}^i}$. Therefore, the response is simulated in such a way that for any target i, $\exists j \in \{1, \ldots, n_c\}$ such that $z_j^i > 0$, and $\sum_{j=1}^{n_c} z_j^i = 1$.

We test the effect of each factor and compute the performance indicator setting $n_c = 20$, $n_t = 22$, $d = 20$, and the proportion of non-zero data $m \in \{0.1, 0.15, 0.2, 0.25\}$. The first (resp. last) half of factors are continuous (resp. discrete) and $\boldsymbol{\beta} \in \mathbb{R}^{d'}$ ($d' = d$ since discrete factors have only two levels) satisfies:

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, -\beta_4, -\beta_5, 0, 0, 0, 0, 0, \beta_1, \beta_2, \beta_3, -\beta_4, -\beta_5, 0, 0, 0, 0, 0) \tag{5}$$

where β_k, $k = \{1, \ldots, 5\}$, are independently drawn from the following uniform distributions:

$$\beta_1, \beta_5 \sim U(8, 10), \quad \beta_2, \beta_4 \sim U(12, 14) \quad \text{and} \quad \beta_3 \sim U(16, 18). \tag{6}$$

The first 10 components of β correspond to the continuous factors, the 10 following components of β correspond to the first level of the discrete factors with two levels (the second level having a null effect). In addition, the standard deviation of the noise is set to $\eta = 10$ that gives a median performance indicator of 0.69 between $\mathbf{z}$ and $\mathbb{X}\boldsymbol{\beta}$ for 100 runs. The main characteristics of the simulation setting are given in Table 2.

In the simulation algorithm proposed above, the link between $\mathbb{X}\boldsymbol{\beta}$ and $\mathbf{z}$ is non-linear due to the step where some values in $\tilde{\mathbf{z}}$ are set to zero (Stage 3 of the algorithm) and the step where $\tilde{\mathbf{z}}$ are scaled to obtain a vector of probabilities for each target (Stage 4). To increase the non-linearity in an additional simulation study, we modify Stage 3 in the algorithm by generating $\tilde{\mathbf{z}}$ in the Gaussian distribution $\mathrm{N}(f(\mathbb{X})\boldsymbol{\beta}, \eta^2\mathbf{I})$, where f transforms four of the continuous factors included in $\mathbb{X}$, two with an expected effect on the response and two without effect given the form of $\boldsymbol{\beta}$ specified above: $f(\mathbb{X}) = (\mathbf{x}_1^2, \exp(\mathbf{x}_2), \sqrt{\mathbf{x}_3}, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6^2, \exp(\mathbf{x}_7), \sqrt{\mathbf{x}_8}, \mathbf{x_9}, \ldots, \mathbf{x_d})$.

3.2 Estimated errors of permutation tests

We assess the performance of the two-tailed permutation test for continuous and discrete factors and different proportions m of non-zero data. Figure 2.a shows the distribution of p-values for each factor for $m = 0.25$ (we get similar results for the other values of m, see Supporting Information, Figure S2). 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. The estimated type I errors of the

Table 2: Main specifications of the simulation study.

Object	Value
Number of Targets (n_t)	22
Number of Contributors (n_c)	20
Number of factors (d)	20
Type of factors	Continuous (10) and discrete (10)
Response variable (Z_j^i)	Simulated proportion
Number of observations ($n_t \times n_c$)	440

test at the risk level 0.05 are given in the top part of Table 3 for different values of m, and show that the test is relatively well calibrated. The type II errors (bottom part of Table 3) are very small for discrete factors whatever the value of m. In contrast, they are larger for continuous factors (in particular those with relatively small effect) and decrease with m. We carried out the same analysis for the one-tailed permutation tests and we obtained very similar results as shown by Supporting Information, Tables S1 to S4.

Remark: Type II errors can be relatively large for continuous factors with weak effect (it is around 0.6 when 10% of response values are non-zero), clearly showing a potential limit of the proposed methodology. However, the large type II errors observed with small m might be an artifact resulting from the simulation scheme. Indeed, by setting to zero the n_0 lowest elements in $\tilde{\mathbf{z}}$ to build $\mathbf{z}$ (see Stage 3 of the simulation algorithm), one simply deletes a part of the information contained in the linear relationship between $\mathbb{X}\boldsymbol{\beta}$ and the initial value of $\tilde{\mathbf{z}}$. In other words, one gets $z_j^i = 0$ for components of $\mathbb{X}\boldsymbol{\beta}$ in a large range of values (i.e., small and intermediate values). In contrast, in real cases, $z_j^i = 0$ means that (i, j) is not likely to be a target-contributor pair, and if $\mathbb{X}\boldsymbol{\beta}$ is consistent with $\mathbf{z}$, its component corresponding to $z_j^i = 0$ should tend be only small, not intermediate, and should therefore reinforce the power of the test. Further investigations are however required to test this artifact assumption.

3.3 Assessment of the 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 2.b 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 its maximum value) approximately when $k = 10$ (vertical line), which corresponds to the actual number of factors having a significant effect. The indicator is robust in the sense that

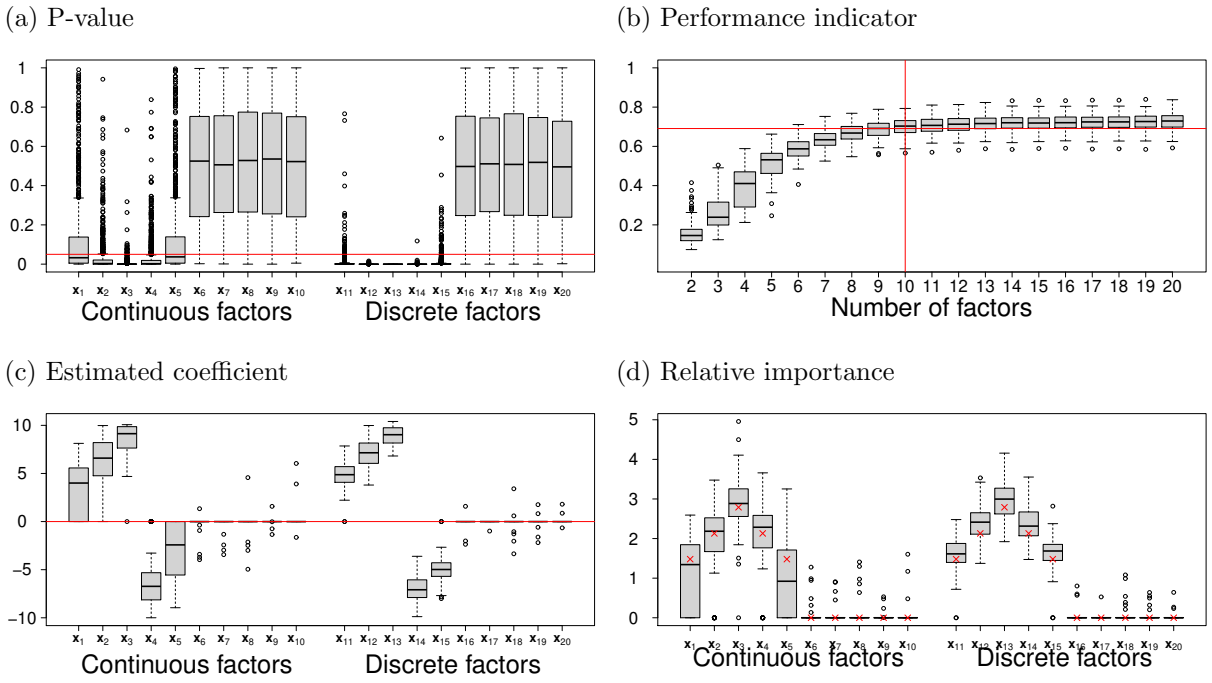


Figure 2: Factor significance and importance in the simulation study. a) P-values of two-tailed permutation tests for each factor (the red line indicates the 0.05 value). b) Distribution of the performance indicator for varying number of included factors (factors are successively incorporated by first including those with lowest p-values). The horizontal red line gives the median value of the performance indicator computed with the true value of $\boldsymbol{\beta}$ (0.69). c) Distribution of estimated coefficients (i.e., the components of $\hat{\boldsymbol{\beta}}$) for each factor. d) Distribution of the relative importance $\tilde{e}_k$ of each factor; the red cross gives the expected relative importance given the simulation scheme described in Section 2.4. The distributions are drawn with $m = 0.25$ and from 1000 repetitions for a) and 100 repetitions for b), c) and d).

adding more factors than the actual number of factors with significant effects does not affect the performance.

3.4 Assessment of the relative importance of factors

Figure 2.c shows that estimated coefficients for non-significant factors are close to zero, while they take values that are clearly positive or negative for significant factors in agreement with the specifications given by Equations (5)–(6). In addition, we note that the amplitudes of the estimated coefficients are not correct, as expected, since the coefficients were constrained between -10 and +10 in the optimization process whereas they were simulated between -14 and +18; see Equations (5)–(6)). However, the relative values of coefficients are estimated

Table 3: Estimated type I and type II errors of the two-tailed permutation tests based on 1000 repetitions in the simulation study (m is the proportion of non-zero values for the response).

Type I	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	0.043	0.044	0.046	0.036	0.047	0.046	0.047	0.049	0.063	0.052
0.15	0.046	0.049	0.051	0.042	0.040	0.054	0.043	0.054	0.054	0.054
0.2	0.057	0.043	0.043	0.056	0.039	0.051	0.034	0.055	0.052	0.045
0.25	0.058	0.043	0.049	0.046	0.050	0.041	0.044	0.055	0.052	0.049

Type II	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	0.607	0.293	0.097	0.321	0.616	0.147	0.008	0.000	0.004	0.159
0.15	0.492	0.214	0.044	0.223	0.522	0.093	0.002	0.000	0.001	0.086
0.2	0.449	0.176	0.037	0.189	0.460	0.063	0.001	0.000	0.001	0.066
0.25	0.433	0.147	0.021	0.151	0.447	0.046	0.000	0.000	0.001	0.041

satisfactorily as deduced from Figure 2.d, which shows that the relative importance $\tilde{e}_k$ of each factor k is approximately distributed around its expected value.

Very similar results are obtained when the standard deviation of the noise η is decreased to $\eta = 5$ or increased to $\eta = 15$, even if the largest standard deviation leads to wider distributions of the relative importance (and higher p-values for factors with non-zero effects, i.e., larger type II errors); see Supporting Information, Figures S3-S4, which are analogous to Figure 2.

Similar results are also obtained when we include additional non-linearity with the transformation function f (see Supporting Information, Figure S5) the sum $\sum_{j=1}^{n_c} z_j^i$ is drawn from a uniform distribution between 0.5 and 1 (which mimics the non observation of some contributors), instead of fixing $\sum_{j=1}^{n_c} z_j^i$ to the value 1 as described in stage 4 of the simulation algorithm detailed in Section 3.1 (see Supporting Information, Figure S6).

In addition, Supporting Information, Figure S7, shows that the estimates of non-zero coefficients are pushed towards the limits of the range, because of the non-identifiability evoked at the end of Section 2.3, when we consider wider constraining intervals for the optimization of $\boldsymbol{\beta}$.

3.5 Assessment of the rank prediction with cross-validation

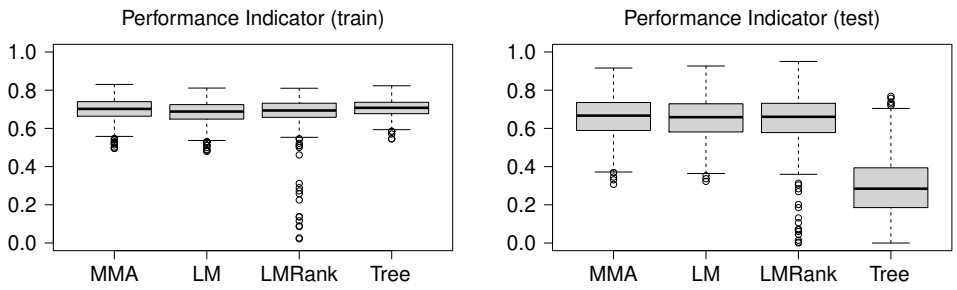
The methodology proposed in this article (MMA) is compared in terms of ranking performance for the simulated data with the three other methods (LM, LMRank and Tree) introduced in Section 2.5. The prediction methods are applied to the set of factors selected by the permutation tests. Figure 3.a shows the good performance of the multitest-based multivariate analysis (MMA). The linear models (LM and LMRank) have relatively similar efficiency, whereas the decision tree (Tree) is clearly less efficient when it is applied to the test samples. Similar results are obtained when one includes the transformation function f to increase the non-linearity (see Figure 3.b), and when one considers the contributor ranking indicator (CR) instead of the performance indicator except for LMRank that is less efficient than MMA and LM based on this criterion (see Supporting Information, Figure S8).

4 Application I: Equine Influenza

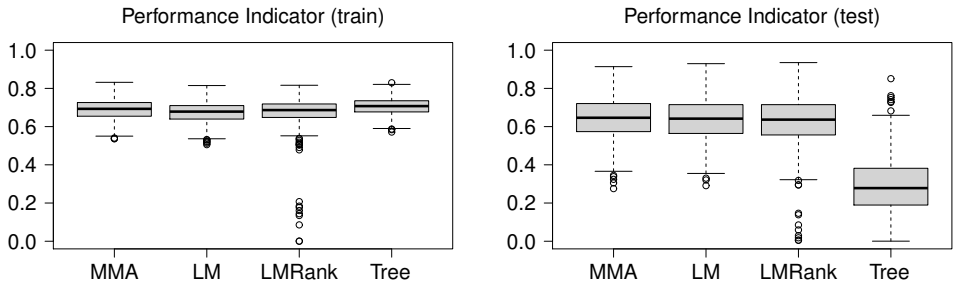
We consider an Equine Influenza outbreak in 2003 in race horses from different training yards in Newmarket. Genomic data collected during this outbreak from 48 horses were studied to explore the virus transmissions across the observed horse population[31]. Intra-host sequences were obtained for each horse and these sequences were used to estimate the probabilities of disease transmission between hosts using the BadTrIP software[32;33,chap. 3]. BadTrip was run in BEAST2[34] using two independent MCMC chains of 5 million steps. We used the dates of first positive swabs as epidemiological data in BadTrIP and allowed the horses to be infected for 9 days except for the first horse A01, which was allowed to be infected for 15 days to provide overlap in the infection periods of the first horses in the transmission chain. The estimated transmission probabilities (shown in Supporting Information, Figure S9) are used in the present study as response proportion data; see Table 4. 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. In what follows, we use four discrete factors and one continuous factor computed from the observed variables 'age', 'sex' and 'training yard' described in Table 5.

Some of these factors are missing for some target-contributor pairs (see Table 5 and Supporting Information, 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. Permutation tests are applied factor by factor on subsets without missing values (if any). The data set used for this study is available on a public archive repository[35]. Remark: A non-completely random structure of missing values could generate biased results, but we did not observe clear signs of such a structure in this case study. We simply observed that the age and sex variables were systematically missing for one of the yards with several infected

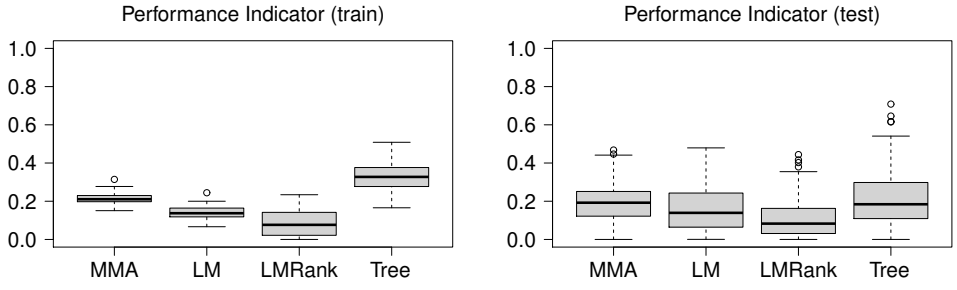
(a)



(b)



(c)



(d)

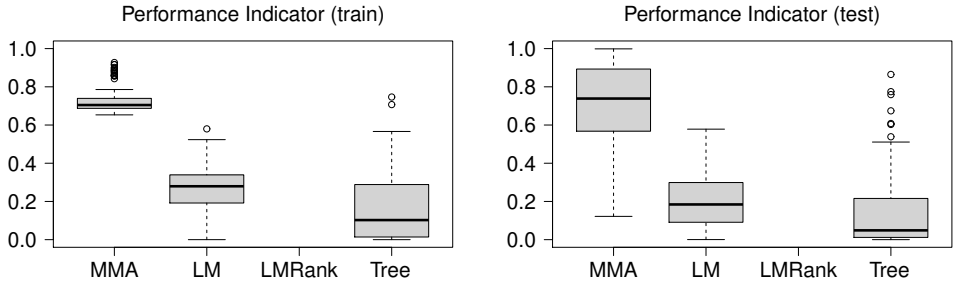


Figure 3: Boxplots of the performance indicator calculated from the training and test samples for MMA, LM, LMRank and Tree: a) in the simulation study without the non-linear transformation f; b) in the simulation study with the non-linear transformation f, c) in the Equine Influenza application; d) in the COVID-19 application (the performance indicator could not be computed for LMRank in this case because this method only predicts null probabilities, which result on a null variance of the ranks and, therefore, an undefined value for the performance indicator).

Table 4: Main specifications of the equine influenza application. Note that the number of observations is $n_t \times (n_c - 1)$ and not $n_t \times n_c$ because we do not account for auto-infection (i.e. for pairs where the contributor-horse is also the target-horse).

Object	Value
Targets/Contributors	Horses
Number of Targets/Contributors (n_t/n_c)	48
Type of factors	Variables related to horses and their environment
Number of factors (d)	5
Response variable (Z_j^i)	Transmission probabilities
Number of observations ($n_t \times (n_c - 1)$ pairs)	2256

Table 5: Explanatory factors for the equine influenza application.

Factor	Description	Number of NA (%)
Same_Yard	1 if the target and contributor are trained in the same yard 0 otherwise	0 (0%)
Same_Sex	1 if the target and contributor have the same sex 0 otherwise	1606 (71%)
Diff_Age	0 if the target and contributor have the same age 1 for a one-year difference 2 for more than one year	1554 (69%)
Dist_Yard	geographic distance (in km) between the training yards of the target and the contributor	0 (0%)
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 "M→M" if a male infected another male	1606 (71%)

horses but nothing else in particular characterizes this yard based on the available data.

Factors Same_Yard, Same_Sex, Dist_Yard and Trans_Sex are significantly correlated to

Table 6: 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

the transmission probability whereas Diff_Age is not; see Table 6. Among these factors, Same_Yard, Same_Sex and two modalities of Trans_Sex have the largest and comparable relative importance with respect to the average factor; see Figure 4. The post-hoc statistic of Same_Yard (T in the bottom part of Table 6) and the Pearson correlation of Dist_Yard (r_s in the top part of Table 6) 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 Table D.1 in Supporting Information, Appendix D.

In addition, there is more intersex transmission when females are the sources (F→M), than when males are the sources (M→F). Supporting Information, Appendix D, shows that the significance of gender-related factors is neither confounded with the effect of the other available factors nor a consequence of heterogeneous sex frequencies.

The performance indicator is calculated on the table containing the four selected factors, discarding the transmissions containing one or more missing values (NA). The performance indicator takes the value $I_{\hat{\boldsymbol{\beta}}}(\mathbb{X}, Z) = 0.21$ 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.

To investigate the robustness and the quality of our approach in this case study, we apply cross-validation and perform the comparison with the three benchmark methods presented in Section 2.5. For this comparison we use the four factors selected with the permutation tests. The decision tree seems to outperform the multivariate analysis on the training samples but this is not confirmed on the test samples; see Figure 3.c. Both methods tend to be more efficient than the two linear models, possibly because of the highly discrete nature of the factors: three factors among four are discrete and the continuous factor (Dist_Yard) takes only 37 different values out of 650 observations. Similar conclusions are drawn when the ranking performance is measured with the CR indicator; see Supporting Information, Figure S10. Remark. In an additional simulation study, we fixed all the coefficients of the continuous factors at zero to see whether the performance of the decision tree is improved when only discrete factors have an impact on the response variable. No significant improvement of the decision tree performance was observed (compare Figure 3.a and Supporting Information, Figure S11) and further investigations are required to understand the reason why the decision tree is relatively efficient in the influenza case study.

5 Application II: COVID-19

A recent article 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[2]. 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/`.

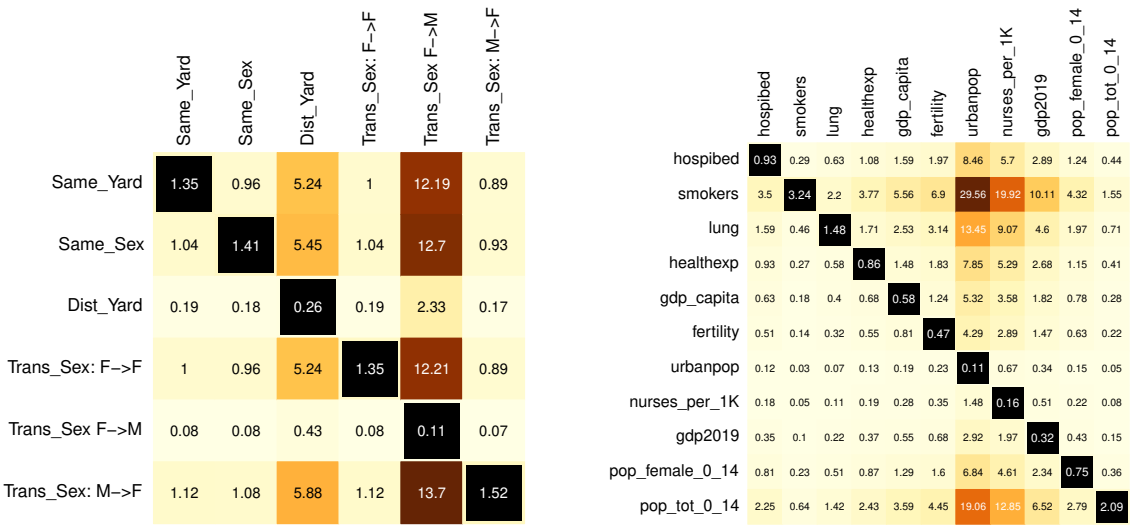


Figure 4: Relative importance of significant factors for the Equine influenza (left) and COVID-19 (right) applications. Diagonal values give the relative importance $\tilde{e}_k$ of each factor k with respect to the average factor as defined by Equation (4). Non-diagonal values give the relative importance of any factor with respect to any other factor, i.e. $\tilde{e}_k/\tilde{e}_{k'} = e_k/e_{k'}$; large values are highlighted with brownish colors.

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 23 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[36] and The Covid Tracking Project (`https://covidtracking.com`). The choice of considering Northern American targets and European contributors was made because Europe was on 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 variables related to economy, demography, health, healthcare system and climate; see Table 7 and Supporting Information, Table S6. 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 variable, the higher the mixture probability. Consequently, we use the univariate test (iii) for continuous factors, which are computed for each target-contributor pair by $x_k^{(i,j)} = |x_k^i - x_k^j|$. The data set used for this study is available on a public archive repository[37].

Table 7: Main specifications of the Covid-19 application.

Object	Value
Targets	States from the USA and provinces from Canada
Number of Targets (n_t)	23
Contributors	Members of the European Economic Area
Number of contributors (n_c)	21
Type of factors	Variables related to economy, demography, health, healthcare system and climate
Number of factors (d)	29
Response variable (Z_j^i)	Mixture probabilities
Number of missing values	0
Number of observations ($n_t \times n_c$)	483

Figure 5 shows the p-values obtained for each factor and the Spearman's correlation for significant factors. We identified eleven impacting factors whose definitions are provided in Supporting Information, Table S6: 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 (since we expect that the similarity of the mortality dynamics of two countries decreases when the difference in the factor values for the two countries increases).

Then, we applied the multivariate analysis based on the eleven significant factors. The factors smokers (percentage of smokers within the population), pop_tot_0_14 (percentage of population in the age group 0-14) and lung (death rate for lung diseases per 100,000 people) have the largest relative importance in the linear combination of factors explaining the mortality dynamics similarity; see Figure 4. Hence, small deviations of these factors strongly favor the similarity of mortality curves. In addition, the performance indicator is equal to $I_{\hat{\boldsymbol{\beta}}}(\mathbb{X}, Z) = 0.73$, which shows that a high monotonous dependency exists between the mixture probabilities and these factors. This is confirmed by Supporting Information, Figure S12, which illustrates the relative good match between predicted ranks based on selected factors and mixture probabilities interpreted as probabilities of similarity.

Figure 3.d and Supporting Information, Figure S13, show the high performance of the MMA with respect to LM and Tree (we could not provide, for this case study, the performance of LMRank as explained in the caption of Figure 3). The performance indicator is quite variable when it is computed from the test samples. In comparison the CR indicator is

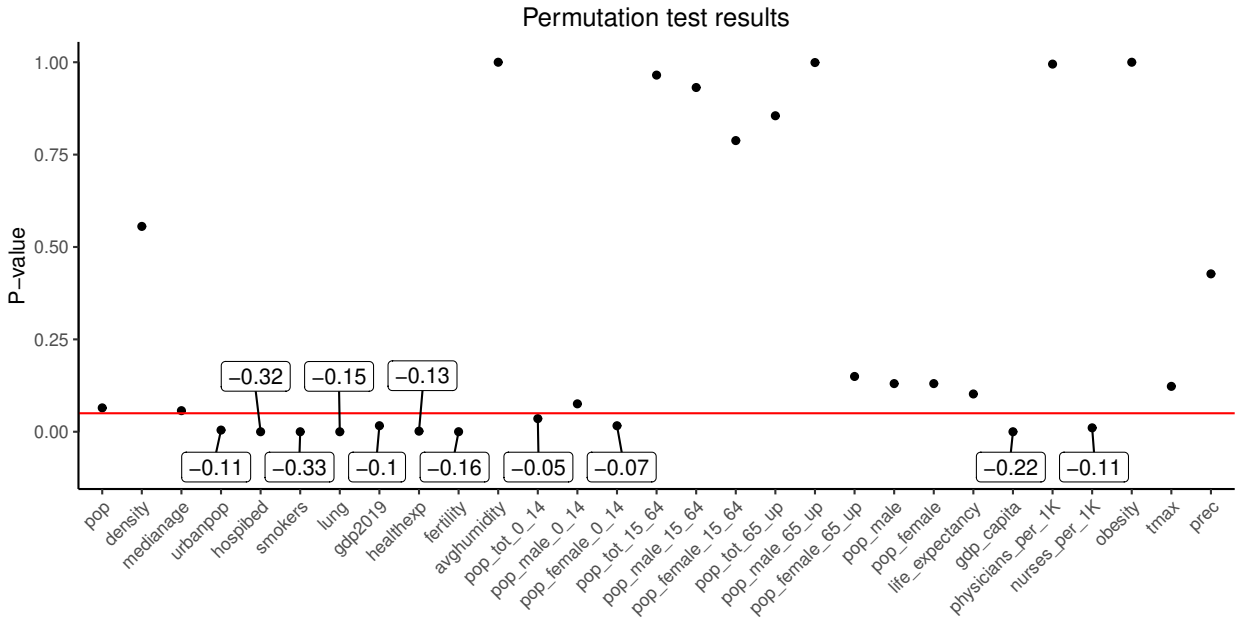


Figure 5: Comparison of the p-values (black dots) for the univariate tests (iii) associated to each factor and the significance level $\alpha = 0.05$ (horizontal line). The Spearman correlation is given by the framed value for factors with a p-value below the threshold.

less variable. However, in both cases, MMA remains globally more efficient than the other methods.

6 Discussion

For exploring the relationships between zero-inflated proportion data (ZIPD) and potentially impacting factors, we developed permutation tests explicitly taking into account the dependence structure in the proportions. These tests being uni-dimensional (i.e., each factor being treated separately), we propose a posterior analysis simultaneously handling the multiple factors. This posterior analysis is grounded on a performance indicator quantifying the percentage of correlation explained by the subset of significant factors and enables the ranking of proportions based on the observation of the factors. Ranking the proportions as we proposed is particularly useful in epidemiology because it especially allows the identification of the most likely sources of infection for a given recipient host simply from host characteristics.

Based on the simulation study, the uni-dimensional tests are well calibrated with respect to type I error in the situations under consideration and, overall, type II error is satisfactory

and generally lower for discrete factors than for continuous factors. These results are consistent with the generally-observed reliability of permutation tests. The main assumption of permutation tests (that are distribution-free) is the exchangeability of data under the null hypothesis. However, this assumption can be violated when the dependence structure is known and especially in the case of block permutations[16]. This structure must be taken into account during permutations otherwise the tests lose power (see Supporting Information, Appendix A). This is why we test the null hypothesis only on a subset of all possible permutations.

The performance indicator, allowing a simultaneous treatment of all the factors, was shown to provide a relevant and parsimonious description of the strength of the link between ZIPD and multiple factors. Indeed, a plateau in terms of performance is reached when all the significant factors are added. We also observed a relative robustness in the performance using training and test sets in a cross-validation framework applied to the simulation and real studies. Cross-validation was also used to compare the performance of our multivariate analysis with respect to the linear regression model, the rank-based linear regression model and the decision tree. Contrasted results were obtained across the three studies, but the multitest-based multivariate analysis appeared as a versatile approach adapted to continuous, discrete and mixed factors.

We have challenged our approach in various settings. However, several particularities of the data not tested in the simulation study could be considered to deepen the conditions of validity of the tests and of the performance indicator in particular (for example, particularities concerning the collinearity or dependence between the factors, the proportion and the structure of missing values, and the number of modalities of discrete factors). The code accompanying this article will facilitate the exploration of the limitations of our approach.

Methodological perspectives Target/contributor pairs containing missing values (`NA`) are handled differently for the permutation tests and for the calculation of the indicator (see application on equine influenza in Section 4). Since the permutation tests are applied factor by factor, the pairs containing a `NA` for a factor are removed only for the test associated to this factor. In contrast, for the calculation of the performance indicator, which simultaneously handles the factors, any pair containing a `NA` for any selected factor is ignored. A more subtle treatment of `NA`, in particular for the calculation of the indicator, deserves to be explored.

If the zero-inflation in the response variable is accounted for in the performance indicator using a scaling term (see Equation (3)), it is not specifically addressed in the permutation tests because these tests are valid despite the zero-inflation and, in practice, work relatively well. Nevertheless, balancing techniques[8] such as those evoked in the introduction might be adapted to the dependency structure[38] and applied to eventually improve the performance

of the approach. Other approaches grounded on distinct treatments of zeros and non-zeros values and a resulting modification of the test statistics may also be considered to explicitly handle the zero-inflation in the permutation test[39,40,41].

A non-negligible computational cost in our approach is due to the optimization of $\boldsymbol{\beta}$ with a genetic algorithm (the dimension of the optimization domain is equal to the number of significant factors retained with the uni-dimensional tests). This cost might be reduced with linear algebra and analysis tools[42].

The method that we propose only provides a point estimate of the performance indicator $I_{\hat{\boldsymbol{\beta}}}(\mathbb{X}, \mathbf{z})$ (that depends on $\mathbb{X}$ and $\hat{\boldsymbol{\beta}}$ via the term $M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$) and point predictors of the ranks of $\mathbf{z}$ (i.e., $\hat{R}_{\mathbf{z}}$) within each target-contributors block using $M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$ (see Sections 2.3–2.5). The robustness of these estimates can be approached by the cross-validation technique that we use in this manuscript. A more advanced evaluation of their uncertainty could be developed in further study by propagating two sources of uncertainty: *(i)* the uncertainty resulting from the test outputs that lead the components of $\boldsymbol{\beta}$ to be set to zero or to be optimized in the maximization of $r_s^2(M_{\mathbb{X}}\boldsymbol{\beta}, \mathbf{z})$, *(ii)* the uncertainty associated to the maximization of $r_s^2(M_{\mathbb{X}}\boldsymbol{\beta}, \mathbf{z})$ given the components of $\boldsymbol{\beta}$ set to zero. If a method can be developed to account for both sources of uncertainty, one would be able to assess the uncertainty of $M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$ and subsequently assess the uncertainty of the performance indicator estimate and the rank predictors. Practically, non-parametric bootstrap may allow us to derive an empirical distribution of $M_{\mathbb{X}}\hat{\boldsymbol{\beta}}$ and, therefore, empirical distributions of $I_{\hat{\boldsymbol{\beta}}}(\mathbb{X}, \mathbf{z})$ and $\hat{R}_{\mathbf{z}}$.

The indicators that we use to evaluate the performance of the method in the simulation study and the applications, namely the performance indicator I and the contributor ranking indicator CR, reflect distinct properties of our approach: I measures the adequateness of the prediction of all proportion ranks whereas CR focuses on the rank prediction for proportions whose actual values are positive. Additional indicators could be envisioned. For instance, one could consider an indicator evaluating how much the target-contributor pair with the largest proportion is correctly ranked for any target (in other words, this indicator would assess the ability of the method(s) to identify, for any target, the most likely target-contributor pair).

Interestingly, the uni-dimensional tests can be applied as a first stage for performing an initial factor selection, whatever the posterior multivariate analysis that is carried out. Here, in addition to our multivariate analysis, we considered relatively simple tools for the posterior analysis, namely linear regression and decision tree. We could consider more complex approaches with known ability to handle, e.g., non-linearity and interactions of high order, and optionally embedding an additional factor-selection stage to solve possible issues generated by eventual dependence between factors selected with our unit tests. Thus, one could explore for example the use of neural networks[43], multivariate adaptive regression splines[44], random forest[45] and boosted generalized linear models[46,47].

In the two applications that we considered, the response variable (i.e., the transmission probability or the similarity probability) is actually estimated from external data and we only consider a point estimate of the probability for each pair. Thus, we deal with the first level of uncertainty, namely the fact that the transmission or the similarity is uncertain and therefore represented by a probability instead of a true/false variable. However, we do not handle the second level of uncertainty, namely the uncertainty of the probability estimates. This source of uncertainty could be handled as follows: Suppose that we have at disposal distributions of probabilities (e.g., posterior distributions obtained from a Bayesian approach) instead of point estimates, then we could propagate the uncertainty about the probabilities[48] into our test by (i) sampling the probabilities from their distributions and (ii) applying the permutation to each sample of probabilities (at stage 2 of the conditional Monte-Carlo algorithm described in Section 2.2). Moreover, by sampling the probabilities from their distributions, we could obtain the distribution of the performance indicator given by Equation (3) instead of a single value. It has however to be noted that if the estimation of $\mathbf{z}$ is biased and hence the weights of network edges are misspecified, the method proposed in this article will certainly miss impacting factors and possibly lead to the identification of unimportant factors as significant.

Epidemiological perspectives The study of Equine Influenza data confirmed the obvious importance of direct contact between hosts for virus transmission: the more frequent contact between horses (same or nearby yard), the higher the probability of transmission. The interpretation of sex differentiation in transmission potential is more complex. We have ruled out an issue of confounding effect between sex and other variables available in the data set and an issue of sex balancing in the observed population of horses. Behavioral, immunological, physiological or organizational factors should be explored to unravel the mechanisms explaining the excess of female-to-female transmissions and transmissions in which a female is the source of infection. Typically, boys were shown to more likely transmit H1N1 to boys and girls to girls probably as a result of assortative mixing among playmates[49]. Another potential explanation could stem from the tendency of horse owners to group horses according to gender, in an attempt to reduce aggressive interactions and the risk of injuries[50]. More generally, given the low value of the performance indicator (0.21), it would be interesting to introduce other factors in the analysis of transmissions by considering other equine influenza data sets, which would allow the results obtained in this article to receive further checks.

Beyond this case study, estimating the probabilities that individuals are linked by transmission events during epidemics of infectious diseases or by progeny relationships in population dynamics has been the subject of numerous studies tackled, in particular, with joint models of epidemiological dynamics and evolutionary processes[1,51,52,53,54,55,56] or with phy-

logeny, phylogeography and some forms of birth–death processes[32,57,58,59,60,61,62]. In many of these studies, it would be interesting to take matters further by exploring the statistical relationship between the inferred links (generally corresponding to ZIPD) and factors characterizing the individuals and the environment. Indeed, determining how factors favor the spread of pathogens or species is crucial to better understand the underlying dynamics[63] and to design adequate control or conservation strategies. In the phylogeography literature, a framework grounded on randomization was proposed to test hypotheses about the effect of environmental variables on pathogen spatial spread[64], but this framework requires the spatio-temporal reconstruction of phylogenetic trees using a software such as BEAST. In contrast, the method that we propose can be applied to random transmission trees and random phylogenetic trees, whatever the way these trees are obtained.

The study of COVID-19 data allowed us to correlate similarity in COVID-19 mortality curves with similarity in certain macroscopic factors related to demography, economy, population health and healthcare system. The high value of the performance indicator (0.73) and its relative robustness observed via cross-validation indicate that these factors can be used to predict with a certain accuracy the similarity between COVID-19 mortality dynamics in geographic entities of Northern America and Europe. Thus, these factors can be viewed as characteristics intrinsically measuring the preparedness and/or vulnerability of geographic entities. The unexplained part of the rank correlation between the mixture probabilities and the linear combinations of factor differences could be due, in particular, to heterogeneous initial conditions of the outbreaks and heterogeneous control measures. It would be interesting to include in the analysis explanatory variables reflecting these components of the epidemics. Furthermore, the analysis that we performed is based on mortality data up to June 6, 2020, which approximately correspond to the first wave of the COVID-19 epidemics in Northern America and Europe. It would be interesting to repeat this analysis across time to assess the temporal (in-)stability of our findings. Specifically, one could expect a change in significant factors between the first and the second (or subsequent) epidemic waves, especially because of heterogeneous levels of immunity after the first wave and heterogeneous impacts of the first wave on the awareness of populations.

Acknowledgements

This work was funded by an ANR grant (SMITID project; ANR-16-CE35-0006).

Data availability statement

The data that support the findings of this study are openly available[35,37] in the public archive repository Zenodo at `http://doi.org/10.5281/zenodo.4837560` and `http://doi.org/10.5281/zenodo.4769671`. `R` codes are available in the package `ZIprop` deposited in `R` CRAN (`https://cran.r-project.org/package=ZIprop`) and GitLab (`https://gitlab.paca.inrae.fr/meribaud/ziprop`). The package also includes the data sets used in this study.

Supplementary Materials

Appendices A to D, Figures S1 to S13 and Tables S1 to S6 cited in the manuscript are available in a single Supporting Information PDF file.

References

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

[2] Soubeyrand S, Ribaud M, Baudrot V, Allard D, Pommeret D, Roques L. COVID-19 mortality dynamics: The future modelled as a (mixture of) past(s). *PLoS ONE* 2020; 15(9): e0238410.

[3] Soubeyrand S, Sache I, Lannou C, Chadœuf J. Residual-based specification of the random-effects distribution for cluster data. *Statistical Methodology* 2006; 3: 464-482.

[4] Soubeyrand S, Chadœuf J. Residual-based specification of a hidden random field included in a hierarchical spatial model. *Computational Statistics & Data Analysis* 2007; 51: 6404–6422.

[5] Weisberg S. *Applied linear regression*. 528. John Wiley & Sons . 2005.

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

[7] Nelder JA, Wedderburn RW. Generalized linear models. *Journal of the Royal Statistical Society: Series A* 1972; 135(3): 370–384.

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

[9] Stasinopoulos DM, Rigby RA, others . Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software* 2007; 23(7): 1–46.

[10] Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C* 2005; 54(3): 507–554.

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

[12] Douma JC, Weedon JT. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Methods in Ecology and Evolution* 2019; 10(9): 1412–1430.

[13] Tang ZZ, Chen G. Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics* 2019; 20(4): 698–713.

[14] Welch WJ. Construction of permutation tests. *Journal of the American Statistical Association* 1990; 85: 693–698.

[15] Duffy DE, Quiroz AJ. A permutation-based algorithm for block clustering. *Journal of Classification* 1991; 8: 65–91.

[16] Winkler AM, Webster MA, Vidaurre D, Nichols TE, Smith SM. Multi-level block permutation. *Neuroimage* 2015; 123: 253–268.

[17] Keele L, Miratrix L. Randomization inference for outcomes with clumping at zero. *The American Statistician* 2019; 73: 141–150.

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

[19] Breiman L, Friedman J, Stone C, Olshen R. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole Statistics-Probability SeriesTaylor & Francis . 1984.

[20] Hollander M, Wolfe DA, Chicken E. *Nonparametric statistical methods*. 751. John Wiley & Sons . 2013.

[21] Kendall MG. The treatment of ties in ranking problems. *Biometrika* 1945; 33: 239–251.

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

[23] Spearman C. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 1904; 15(1): 72–101.

[24] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 1952; 47(260): 583–621.

[25] Dunn OJ. Multiple comparisons using rank sums. *Technometrics* 1964; 6(3): 241–252.

[26] Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; 75: 800–802.

[27] Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Statistics in Medicine* 1990; 9: 811–818.

[28] Mebane Jr WR, Sekhon JS. Genetic optimization using derivatives: the rgenoud package for R. *Journal of Statistical Software* 2011; 42(11): 1–26.

[29] Kloke JD, McKean JW. Rfit: Rank-based estimation for linear models. *The R Journal* 2012; 4(2): 57–64.

[30] Jaeckel LA. Estimating regression coefficients by minimizing the dispersion of the residuals. *The Annals of Mathematical Statistics* 1972; 43(5): 1449–1458.

[31] Hughes J, Allen RC, Baguelin M, et al. Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. *PLoS Pathogens* 2012; 8(12): e1003081.

[32] De Maio N, Worby CJ, Wilson DJ, Stoesser N. Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS Computational Biology* 2018; 14(4): e1006117.

[33] Alamil M. *Reconstruction of the transmission of a virus during an epidemic by statistical learning on genomic data*. PhD thesis. Aix-Marseille University, Marseille; 2020.

[34] Bouckaert R, Heled J, Kühnert D, et al. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 2014; 10: e1003537.

[35] Ribaud M, Hughes J. Equine Influenza dataset. *10.5281/zenodo.4837560* 2021. doi: 10.5281/zenodo.4837560

[36] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* 2020; 20: 533–534. doi: https://doi.org/10.1016/S1473-3099(20)30120-1

[37] Ribaud M, Martinetti D, Soubeyrand S. Data for the comparison of COVID-19 mortality in European and North American geographic entities. *10.5281/zenodo.4769671* 2021. doi: 10.5281/zenodo.4769671

[38] Lahiri S. *Resampling methods for dependent data*. New York: Springer-Verlag . 2003.

[39] Hallstrom AP. A modified Wilcoxon test for non-negative distributions with a clump of zeros. *Statistics in Medicine* 2010; 29: 391–400.

[40] Pimentel RS, Niewiadomska-Bugaj M, Wang JC. Association of zero-inflated continuous variables. *Statistics & Probability Letters* 2015; 96: 61–67.

[41] Finos L, Pesarin F. On zero-inflated permutation testing and some related problems. *Statistical Papers* 2020; 61: 2157–2174.

[42] Alfons A, Croux C, Filzmoser P. Robust maximum association between data sets: The R package ccaPP. *Austrian Journal of Statistics* 2016; 45: 71–79.

[43] Warner B, Misra M. Understanding neural networks as statistical tools. *The American Statistician* 1996; 50: 284–293.

[44] Friedman JH. Multivariate adaptive regression splines. *The Annals of Statistics* 1991; 19: 1–67.

[45] Breiman L. Random forests. *Machine Learning* 2001; 45: 5–32.

[46] Friedman JH. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 2002; 38: 367–378.

[47] Bühlmann P, Yu B. Boosting with the L 2 loss: regression and classification. *Journal of the American Statistical Association* 2003; 98: 324–339.

[48] Georgescu V, Soubeyrand S, Kretzschmar A, Laine AL. Exploring spatial and multitype assemblages of species abundances. *Biometrical Journal* 2009; 51: 979–995.

[49] Cauchemez S, Bhattarai A, Marchbanks TL, et al. Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proceedings of the National Academy of Sciences* 2011; 108: 2825–2830.

[50] Jørgensen GHM, Borsheim L, Mejdell CM, Søndergaard E, Bøe KE. Grouping horses according to gender—effects on aggression, spacing and injuries. *Applied Animal Behaviour Science* 2009; 120: 94–99.

[51] Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Computational Biology* 2014; 10: e1003457.

[52] Lau MS, Marion G, Streftaris G, Gibson G. A systematic Bayesian integration of epidemiological and genetic data. *PLoS Computational Biology* 2015; 11: e1004633.

[53] Lau MS, Gibson GJ, Adrakey H, et al. A mechanistic spatio-temporal framework for modelling individual-to-individual transmission—With an application to the 2014-2015 West Africa Ebola outbreak. *PLoS Computational Biology* 2017; 13: e1005798.

[54] Mollentze N, Nel LH, Townsend S, et al. A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proceedings of the Royal Society B* 2014; 281: 20133251.

[55] Morelli MJ, Thébaud G, Chadœ uf J, King DP, Haydon DT, Soubeyrand S. A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Computation Biology* 2012; 8: e1002768.

[56] Soubeyrand S. Construction of semi-Markov genetic-space-time SEIR models and inference. *Journal de la Société Française de Statistique* 2016; 157(1): 129–152.

[57] Didelot X, Fraser C, Gardy J, Colijn C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution* 2017; 34(4): 997–1007.

[58] Hall M, Woolhouse M, Rambaut A. Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set. *PLoS Computational Biology* 2015; 11: e1004613.

[59] Leitner T, Romero-Severson E. Phylogenetic patterns recover known HIV epidemiological relationships and reveal common transmission of multiple variants. *Nature Microbiology* 2018; 3: 983.

[60] Pybus OG, Suchard MA, Lemey P, et al. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences* 2012; 109: 15066–15071.

[61] Rakotomalala M, Vrancken B, Pinel-Galzi A, et al. Comparing patterns and scales of plant virus phylogeography: Rice yellow mottle virus in Madagascar and in continental Africa. *Virus Evolution* 2019; 5: vez023.

[62] Valdazo-González B, Kim JT, Soubeyrand S, et al. The impact of within-herd genetic variation upon inferred transmission trees for foot-and-mouth disease virus. *Infection, Genetics and Evolution* 2015; 32: 440–448.

[63] Picard C, Dallot S, Brunker K, et al. Exploiting genetic information to trace plant virus dispersal in landscapes. *Annual Review of Phytopathology* 2017; 55: 139–160.

[64] Dellicour S, Rose R, Pybus OG. Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinformatics* 2016; 17: 1–12.