Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-024-02715-1

ORIGINAL PAPER

Some combinatorics of data leakage induced by clusters

Fabian Guignard[1] · **David Ginsbourger**[1] · **Lilia Levy Häner**[2] · **Juan Manuel Herrera**[2]

Accepted: 25 March 2024

Abstract
Data leakage is a common issue that can lead to misleading generalisation error estimation and incorrect hyperparameter tuning. However, its mechanisms are not always well understood. In this work, we consider the case of clustered data and investigate the distribution of the number of elements in leakage when the data set is uniformly split. For both the validation and test sets, the first and second moments of the number of elements in leakage are derived analytically. Modelling consequences are investigated and exemplified on simulated data. In addition, the case of an actual agronomic feasibility study is presented. We demonstrate how data leakage can distort model performance estimation when an inadequate data splitting strategy is used. We provide an understanding of data leakage in the context of clustered data by quantifying its role in predictive modelling. This sheds light on related challenges that may impact the practice in agronomy and beyond.

Keywords Group leakage · Clustered data · Dependent data · Predictive modelling · Prediction error estimation · Generalisation error estimation

1 Introduction

Model selection and assessment through validation procedures are crucial in statistical modelling and machine learning. Estimating a given algorithm's generalisation performance helps to select the learning method, choose its right flexibility level, and quantify its predictive quality (Hastie et al. 2009). However, in situations where the data are structured in groups or clusters, baseline validation procedures may lead to misleading estimations of the generalization error. Clustered/grouped data are ubiquitous in agronomy and beyond, where they can for instance stand for observations at a neighbouring time, at neighbouring locations, when using the same crop variety, or in longitudinal studies where several observations are made on the same individual (e.g., in health sciences). Depending on how the clusters are shared between the validation/testing set and the learning set, they can foster undesired information exchange between the different subsets, a phenomenon known as data leakage (Kaufman et al. 2012; Kapoor and Narayanan 2023). Therefore, the algorithm visiting a cluster during the learning phase takes advantage of this knowledge when it is used on data kept for the generalization error estimation that belong to the same cluster. This may lead to the selection of inadequate models and hyper-parameters, notably when using distance methods such as nearest neighbours prediction.

To illustrate this phenomenon, let us consider the following practical situation, which is one of the cases that will be presented, motivated and discussed in detail in Section 4. An applied scientist or an analytics consultant is commissioned to conduct a feasibility study in the field of agricultural sciences. The aim of the considered study is to determine whether the yield of winter wheat can be predicted based on the wheat variety and its environment given a specific region. The yield of six winter wheat varieties is measured in three replicates under 23 different environmental conditions, totalling 414 observations. The analyst uses a

✉ Fabian Guignard
fabian.guignard@unibe.ch

David Ginsbourger
david.ginsbourger@unibe.ch

Lilia Levy Häner
lilia.levy@agroscope.admin.ch

Juan Manuel Herrera
juan.herrera@agroscope.admin.ch

[1] Institute of Mathematical Statistics and Actuarial Science, University of Bern, Alpeneggstrasse 22, Bern CH-3012, Switzerland

[2] Plant-Production Systems, Agroscope, Route de Duillier 50, Nyon CH-1260, Switzerland

Published online: 11 April 2024

k-nearest neighbours (KNN) algorithm to predict the respective yields that would be with obtained with each variety in some unobserved environmental condition. The optimal number of neighbours is selected with a 5-fold cross-validation procedure, which provides $k = 1$. To assess the final model, and in accordance with the predictive goal, the same experimental design at 6 unobserved environmental conditions was kept apart as a test set. The mean squared errors are reported in Table 1. As a matter of fact, the test error of the model is far worse than if the analyst used the average response calculated from the learning set. This clear case of overfitting is caused by the clustered nature of the observations, which allows information leakage when the data are not adequately split during the model selection.

In predictive modelling, leakage in a broad sense has been known for a long time (Nisbet et al. 2009; Kuhn and Johnson 2019), and a formal definition was proposed (Kaufman et al. 2012). Data leakage was recently recognized as an important and frequent factor of irreproducibility in machine learning based science (Kapoor and Narayanan 2023). Nevertheless, the mechanisms of data leakage remain formally under-investigated. A common source of data leakage comes from not taking into account the dependencies between the split data sets while these dependencies underlie the predictive goal (Roberts et al. 2017). Indeed, data splitting aims to mimic the relationship between the available data and the set to be generalized. If there is a mismatch between the learning situations encountered within the learning set (by cross-validation) and the out-of-sample prediction situations, then this can lead to unsuitable predictions and quantifications of prediction uncertainty (Kapoor and Narayanan 2023). Several specific cases of data leakage can be seen as special cases of leakage induced by clustered data, such as, from general to particular:

- Leakage induced by hierarchical data (e.g., blocked or nested experimental design), (Roberts et al. 2017);
- Group-structured data, sometimes referred to as group leakage (Ayotte 2021; Ayotte et al. 2021; Meghnoudj et al. 2023). In this case, samples come from the same unit, block, individual, or group;
- Experiment replicates. In various fields, measurements from replicated experiments can be considered a particular case of clustered data, as our agronomical case study illustrates. Let us remark that the response differs for two replicates, while experimental conditions are the same;
- Duplicates in the data set. In this extreme case, some input/outputs pairs are (unintentionally) copied once or several times into the database (Kapoor and Narayanan 2023).

Table 1 Error comparison for the introductory agronomic example. The metric used is the mean squared error

Model	Val. MSE	Test MSE
1-nearest neighbour	89.44	635.54
Avg. of the learning set	–	304.07

In the context of multiple-fold cross-validation (Stone 1974), different solutions have been proposed to account for dependencies within splitting, e.g., by defining fold splitting schemes that mimic the predictive goal (Rice and Silverman 1991) or that are specific to a given problem (Roberts et al. 2017; Montesinos López et al. 2022; Buntaran et al. 2019). Recently, it was proposed to correct the bias of the prediction error estimation obtained by multiple fold cross-validation by taking covariances between the response values into account (Rabinowicz and Rosset 2020).

In many practical situations, the data set is partitioned into three subsets: the learning set on which the model will be fitted, the validation set used to optimize hyperparameters and select the model, and the testing set used to estimate the generalization error and assess the selected model. In this framework, little is known about how the clusters are distributed across the different data subsets.

In the present paper, we analytically investigate the cluster breakdown into the learning, validation and testing sets, aiming at quantifying the data leakage between them and helping to understand its impact on predictive performance. More specifically, the distribution of the number of elements in leakage is characterised, and its first and second moments are derived analytically for both validation and test sets. The impact is illustrated through the analysis of both a synthetic dataset and the agronomical research dataset mentioned above. Overall, the paper aims to clarify the leakage effect of clustered data in practical predictive modelling situations and to raise awareness of some related pitfalls. Although an agronomic application inspires our investigations, the findings of this study could virtually apply to any application domain with clustered data where leakage is likely to occur.

The present article is organized as follows. Starting from general assumptions, probabilistic formulae are derived for quantifying the cluster breakdown between data subsets randomly split in Section 2. In Section 3, the formulae are checked by simulation and are put into practice in a synthetic predictive experiment. Real-world applications on agricultural data are reported in Section 4, where more details on the introductory example are provided and cluster impact under different predictive goals is investigated. Section 5 concludes the paper.

2 Probabilistic modelling

In this section, we characterise the distribution of the number of elements in leakage and derive analytical formulae for resulting expectation and variance. The single-split case is first discussed, followed by the double-split case. Discussion of the obtained formulae give us insight into the data leakage phenomenon when the data set is uniformly split into subsets of prescribed sizes.

2.1 Single-split situation

We consider a finite set $\mathcal{D}$, standing for a clustered data set to be split into disjoint subsets, respectively devoted to training, testing, and also possibly to model validation. We assume furthermore that $\mathcal{D}$ can be partitioned in $n_c \geq 2$ clusters $C_i \subset \mathcal{D}$ $(i = 1, \ldots, n_c)$, i.e. $\mathcal{D} = \bigsqcup_{i=1}^{n_c} C_i$, where $\bigsqcup$ denotes the disjoint union. We denote by $r_i \geq 1$ the size of the cluster C_i, and by $n = \sum_{i=1}^{n_c} r_i$ the cardinality of the data set $\mathcal{D}$.

We first consider the case where $\mathcal{D}$ is randomly split into *learning set* L of prescribed size, used for model training purpose, and a *testing set* $T = L^c$, kept to evaluate prediction errors. Here, L^c denotes the complement of the set L, i.e. the set of elements of $\mathcal{D}$ not in L.

We say that cluster C_i is *in leakage* if $\#(C_i \cap L)\#(C_i \cap T) > 0$ where # is used as symbol of cardinality, that is, if $C_i \cap L \neq \varnothing$ and $C_i \cap T \neq \varnothing$.

Let us remark that the partitions of $\mathcal{D}$ into training/testing subsets and into clusters induce a finer partition as follows:

$$\mathcal{D} = L \sqcup T = \left(\bigsqcup_{i=1}^{n_c} (C_i \cap L) \right) \sqcup \left(\bigsqcup_{i=1}^{n_c} (C_i \cap T) \right).$$

In particular, each cluster is potentially in leakage depending on the split. If it is, information may be shared between the training set and the testing set. Figure 1 provides a sketch of this situation for $n_c = 4$ clusters.

For all $i = 1, \ldots, n_c$, we set $X_i = \#(C_i \cap T)$, which counts the number of elements of a given cluster within the testing set, and denote $X = (X_1, \ldots, X_{n_c})^T$. One important note for the following is that C_i is in leakage if and only if $0 < X_i < r_i$. When a cluster is in leakage, its elements that belong to T will be called *leakage elements*, or simply *leakages*. We define

$$N_i := X_i \, \mathbb{1}_{\{0 < X_i < r_i\}} = X_i \, \mathbb{1}_{\{0 \leq X_i < r_i\}},$$

where $\mathbb{1}$ is the indicator function. For any $i = 1, \ldots, n_c$, the random variable N_i counts the number of leakage elements of cluster C_i. We are interested in counting the total number of leakages, that is

$$N := \sum_{i=1}^{n_c} N_i.$$

As an example, for Fig. 1, we have $X = (0, 5, 1, 5)^T$ and $N = 6$, where the leakages are symbolized by red crosses.

We will now investigate the distribution of N when the testing set is uniformly drawn from $\{T \subset \mathcal{D} \mid \#T = k\}$, where $k < n$ is a prescribed testing set size. First we give the distribution of X.

Theorem 1 *If T is uniformly distributed among subsets of k elements of $\mathcal{D}$, then the random vector X is multivariate hypergeometric distributed with parameters $(k; r_1, r_2 \ldots, r_{n_c})$.*

Proof The proof consists in identifying the data splitting to a sampling without replacement of k marbles from an urn containing n marbles, of which r_1 are of colour C_1, r_2 are of colour C_2, ..., and r_{n_c} are of colour C_{n_c}. There are $\binom{n}{k}$ possibilities of drawing the marbles, among which $\prod_{i=1}^{n_c} \binom{r_i}{x_i}$ contain x_1 marbles of colour C_1, x_2 marbles of colour C_2, ..., x_{n_c} marbles of colour C_{n_c}. Hence, the probability mass function of $X = (X_1, \ldots, X_{n_c})$ is

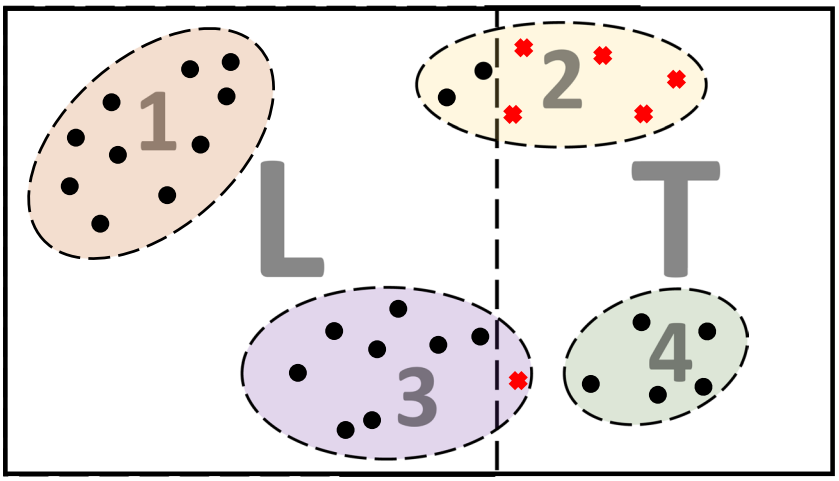

Fig. 1 Sketch of a clustered data set in a single-split situation. Clusters 2 and 3 are in leakage, while clusters 1 and 4 are not. Red crosses represent *leakage elements*, while black points represent the remaining elements

$$\mathbb{P}[X = x] = \frac{\prod_{i=1}^{n_c} \binom{r_i}{x_i}}{\binom{n}{k}},$$

with $x = (x_1, \ldots, x_{n_c})^T$ such that $x_i \in \{0, 1, \ldots, r_i\}$, $i = 1, 2, \ldots, n_c$ and $\sum_{i=1}^{n_c} x_i = k$. □

While the multinomial distribution is obtained by sampling *with replacement* a population with $n_c \geq 2$ categories (binomial distribution in case $n_c = 2$), the multivariate hypergeometric distribution describe its *without replacement* counterpart, and there is a similar relationship between these two distributions as between the binomial and hypergeometric distributions (Johnson et al. 1997). For convenience, we will denote $X \sim \text{Mult. Hypg.}(k; r_1, r_2, \ldots, r_{n_c})$ when the random vector X follows a multivariate hypergeometric distribution with parameters $(k; r_1, r_2, \ldots, r_{n_c})$. For $n_c = 2$, the multivariate hypergeometric distribution reduces to the hypergeometric distribution, that will be denoted $\text{Hypg.}(k; r, n)$. Various properties of the multivariate hypergeometric distribution can be found in Johnson et al. (1997); some that will be useful to us are listed here.

Selected properties of the hypergeometric distribution Let $X = (X_1, \ldots, X_{n_c}) \sim \text{Mult. Hypg.}(k; r_1, r_2, \ldots, r_{n_c})$. Then the following holds:

1. The marginal distributions are hypergeometric, that is, $X_i \sim \text{Hypg.}(k; r_i, n)$, for $i = 1, 2, \ldots, n_c$. In particular,

   $$\mathbb{E}[X_i] = \frac{k}{n} r_i,$$

   and

   $$\text{Var}[X_i] = \frac{k(n-k)}{n^2(n-1)} r_i (n - r_i),$$

   for $i = 1, 2, \ldots, n_c$.
2. The covariance between two components X_i and X_j is given by

   $$\text{Cov}[X_i, X_j] = -\frac{k(n-k)}{n^2(n-1)} r_i r_j,$$

   for $i, j = 1, 2, \ldots, n_c$ where $i \neq j$.
3. For $s \in \{2, \ldots, n_c - 1\}$, the joint distribution of the random vector $(X_{i_1}, X_{i_2}, \ldots, X_{i_s}, k - \sum_{j=1}^{s} X_{i_j})^T$ is

   $$\text{Mult. Hypg.}\left(k; r_{i_1}, r_{i_2}, \ldots, r_{i_s}, n - \sum_{j=1}^{s} r_{i_s}\right).$$

The total number of leakages N is a function of X. As a consequence of Theorem 1 and the properties above, we now derive several results regarding on the distribution of N. The first one concerns the expectation of N.

Corollary 1 *The expected number of leakage elements is given by*

$$\mathbb{E}[N] = k \cdot \left(1 - \frac{1}{n} \sum_{i: r_i \leq k} r_i \frac{\binom{n - r_i}{n - k}}{\binom{n-1}{n-k}}\right). \tag{1}$$

Proof We first remark that for $i = 1, \ldots, n_c$, $X_i = N_i + X_i \mathbb{1}_{\{X_i = r_i\}}$. Summing over all clusters, we get

$$N = k - \sum_{i=1}^{n_c} X_i \mathbb{1}_{\{X_i = r_i\}}, \tag{2}$$

from which, by taking expectations, one obtains $\mathbb{E}[N] = k - \sum_{i=1}^{n_c} r_i \mathbb{P}[X_i = r_i]$. One concludes by evaluating the $\mathbb{P}[X_i = r_i]$ terms using the first of the basic properties of the multivariate hypergeometric distribution listed above and remarking that this probability is null when $r_i > k$. □

Remark By expanding the definition of the binomial coefficient, Eq. 1 can be reformulated in a somehow more straightforward way to make it easier to implement. More precisely, by defining $\binom{n}{k} = 0$ for any integer $n \geq 0$ and any integer k such that $k < 0$ or $k > n$, Eq. 1 can be rewritten as

$$\mathbb{E}[N] = k \cdot \left(1 - \frac{1}{n} \sum_{i=1}^{n_c} r_i \frac{\binom{n - r_i}{n - k}}{\binom{n-1}{n-k}}\right), \tag{3}$$

where the summation index run from 1 to n_c. For the sake of readability, we keep this convention in most of the remainder of the paper.

Let us stress that the second factor of Eq. 3 represents the expected proportion of leakage elements. For $n = 126$, this expected proportion of leakage elements in the testing set is represented in Fig. 2 (various line types, in black) against the split proportion of testing data, k/n, assuming clusters of homogeneous sizes $r = 2, 3, 4, 5, 10, 30$. For a given split proportion, the case $r = 2$ shows the least expected proportion of leakages. The more the cluster size increases, the more the curve bends to the top right corner. For a common splitting proportion of $k/n = 20\%$ (vertical dotted line), most of the testing set contains data for which the corresponding cluster has been visited during the learning phase, even with clusters of small size.

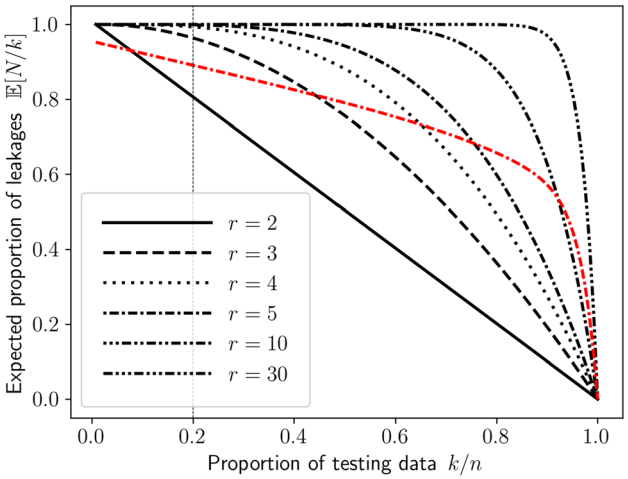

Fig. 2 Expected proportion of leakage elements included in T . In black, $\mathbb{E}[N/k]$ for constant cluster size $r = 2, 3, 4, 5, 10, 30$. In red (dash-dash-dotted line) , $\mathbb{E}[N/k]$ for 6 single individuals, 20 clusters of size 2, 2 clusters of size 5, 1 cluster of size 10 and 2 clusters of size 30

In red (dash-dash-dotted line), $\mathbb{E}[N/k]$ is shown when cluster size varies within $\mathcal{D}$, with $r_1 = \cdots = r_6 = 1, r_7 = \cdots = r_{26} = 2, r_{27} = r_{28} = 5, r_{29} = 10, r_{30} = r_{31} = 30$, which correspond to a data set of size $n = 126$ with 2 clusters of size 30, 1 cluster of size 10, 2 clusters of size 5, 20 pairs and 6 single individuals. Let us remark that the 6 single individuals have the effect of making the red curve start at $\mathbb{E}[N/k] = \frac{120}{126} = \frac{20}{21} = 0.952$, for $k = 1$. The elbow at $k/n = 0.9$ originates from the 2 clusters of size 30, representing almost half of the data.

The next Corollary tells us about the fluctuations of N around its expectation.

Corollary 2

$$\mathrm{Var}[N] = \frac{k}{n}\left(\sum_{i=1}^{n_c} r_i^2 \frac{\binom{n-r_i}{n-k}}{\binom{n-1}{n-k}} + \sum_{i\neq j} r_i r_j \frac{\binom{n-r_i-r_j}{n-k}}{\binom{n-1}{n-k}}\right) - \left(\frac{k}{n}\sum_{i=1}^{n_c} r_i \frac{\binom{n-r_i}{n-k}}{\binom{n-1}{n-k}}\right)^2. \tag{4}$$

Proof Taking the variance of Eq. 2, one has

$$\begin{aligned}
\mathrm{Var}[N] &= \sum_{i=1}^{n_c}\sum_{j=1}^{n_c} \mathrm{Cov}\left[X_i \mathbb{1}_{\{X_i=r_i\}}, X_j \mathbb{1}_{\{X_j=r_j\}}\right] \\
&= \sum_{i=1}^{n_c}\sum_{j=1}^{n_c} r_i r_j \left(\mathbb{P}[X_i = r_i, X_j = r_j] - \mathbb{P}[X_i = r_i]\,\mathbb{P}[X_j = r_j]\right) \\
&= \sum_{i=1}^{n_c} r_i^2\, \mathbb{P}[X_i = r_i] + \sum_{i\neq j} r_i r_j\, \mathbb{P}[X_i = r_i, X_j = r_j] \\
&\quad - \left(\sum_{i=1}^{n_c} r_i\, \mathbb{P}[X_i = r_i]\right)^2.
\end{aligned}$$

The result is obtained by applying the first and third of the properties listed for the multivariate hypergeometric distribution, and by rearranging binomial coefficients. □

Finally, we can get insight into the probability of not observing any leakage between the learning and testing sets.

Corollary 3 *The probability that there is no leakage at all has the following upper bound* :

$$\mathbb{P}[N = 0] \leq \frac{2^{n_c}}{\binom{n}{k}}.$$

Proof No leakage is observed if and only if for all $i = 1, \ldots, n_c$, $X_i \in \{0, r_i\}$ such that $\sum_{i=1}^{n_c} X_i = k$. Hence, the probability of having no leakage is

$$\sum_{x_i\in\{0,r_i\} \,:\, \sum_{i=1}^{n_c} x_i = k} \mathbb{P}[X = x] \leq \sum_{x_i\in\{0,r_i\}} \mathbb{P}[X = x] \leq \sum_{x_i\in\{0,r_i\}} \frac{1}{\binom{n}{k}}.$$

The proof is concluded by noting that there are 2^{n_c} terms in the last sum. □

Remark The probability of no leakage is zero if and only if the testing set size cannot be expressed as a sum of the cluster cardinalities. To prove it, one direction is immediate: from the proof of Corollary 3, if k is not equal to any sum of r_i, then $P[N = 0] = 0$. Conversely, assume without loss of generality that there exists an integer $q \in \{1, \ldots, n_c\}$ such that $k = \sum_{i=1}^{q} r_i$. If $T = \bigsqcup_{i=1}^{q} C_i$ is drawn from among the $\binom{n}{k}$ possible draws, then there is no leakage. Hence,

$$\mathbb{P}[N = 0] \geq \mathbb{P}\left[T = \bigsqcup_{i=1}^{q} C_i\right] = \frac{1}{\binom{n}{k}} > 0.$$

2.2 Double-split situation

In practice, it is common to split $\mathcal{D}$ into three non-empty disjoint subsets: a learning set L, a validation set V, and a testing set T. At first, the model is trained on L, and V is used as an intermediate test set for model selection. Next, the training is performed on $L \cup V$, and $T = (L \cup V)^c$ is used for testing. Thus, all the quantities defined and the results proved in

Subsection 2.1 remain valid for T, because they are not influenced by the fact that the model will be trained on $L \cup V$ once the model selection has been made. Figure 3 shows an analogous sketch as Fig. 1. The same data set is considered, but it is split into three parts this time.

We are now interested in the leakage between the validation set and the learning set. The i-th cluster is in *validation leakage* if $C_i \cap L \neq \emptyset$ and $C_i \cap V \neq \emptyset$. We define the random variable that counts the number of elements belonging to the validation set for a given cluster, $Y_i = \#(C_i \cap V)$, and the random vector $Y = (Y_1, \dots, Y_{n_c})^T$. Describing the number of elements belonging to the learning set for a given cluster will be helpful, so we also define $Z = (Z_1, \dots, Z_{n_c})^T$ with $Z_i = \#(C_i \cap L)$. Note that

$$X_i + Y_i + Z_i = r_i, \qquad i = 1, \dots, n_c.$$

If a cluster is in validation leakage, its elements that belong in V are called *validation leakages*. It is easy to check that C_i is in validation leakage if and only if $0 < Z_i < r_i$ and $0 < Y_i < r_i$. Therefore, a natural definition for counting the number of validation leakages of cluster C_i is

$$\begin{aligned} M_i &:= Y_i \mathbb{1}_{\{0<Y_i<r_i\} \cap \{0<Z_i<r_i\}} \\ &= Y_i \mathbb{1}_{\{0<Y_i \leq r_i\}} \mathbb{1}_{\{0<Z_i \leq r_i\}} \\ &= Y_i \mathbb{1}_{\{0<Z_i \leq r_i\}} \end{aligned}$$

and the total number of validation leakages is

$$M := \sum_{i=1}^{n_c} M_i.$$

As an example, validation leakages are reported in Fig. 3. To avoid confusion, instead of talking about leakage as defined in Subsection 2.1, we will speak of *testing leakage* when considering a leakage between $L \cup V$ and T.

Theorem 2 *Assume that (T, V, L) is uniformly distributed among partitions of $\mathcal{D}$ such that T, V, L have k, ℓ, m elements, respectively (with $k + \ell + m = n$). Then, X, Y, Z and each of these random vectors conditioned by another are multivariate hypergeometric distributed. In particular,*

$$Z \sim \text{Mult.Hypg.}(m; r_1, r_2, \dots, r_{n_c}), \text{ and}$$

$$Y \mid Z = z \sim \text{Mult.Hypg.}(\ell; r_1 - z_1, r_2 - z_2, \dots, r_{n_c} - z_{n_c}),$$

$$\text{for } z = (z_1, \dots, z_{n_c})^T \text{ such that } z_i \in \{0, 1, \dots, r_i\},$$
$$i = 1, 2, \dots, n_c \text{ and } \textstyle\sum_{i=1}^{n_c} z_i = m.$$

Proof The fact that the three vectors X, Y, Z considered separately are multivariate hypergeometric distributed can be seen as a consequence of Theorem 1 given that T, V, L considered separately are uniformly distributed among subsets of $\mathcal{D}$ with respective cardinalities k, ℓ, m. One can in fact check that constructing (T, V, L) iteratively by drawing each subset uniformly (among subsets of relevant cardinalities) in the complement of the set of elements drawn so far, the triplet follows indeed the targeted uniform distribution. The statement about conditional distributions (say of $Y \mid Z = z$) follows from a related argument. In fact, knowing that $Z = z$ informs us in two respects. First, it tells us what L is (and does not inform us further on how L^c is partitioned into T and V), so we are back to the settings of Theorem 1 when it comes to the conditional distribution of V given L. Second, knowing $Z = z$ updates the number of V elements that can be sampled from C_i from r_i to $r_i - z_i$ $(i = 1, \dots .n_c)$ so

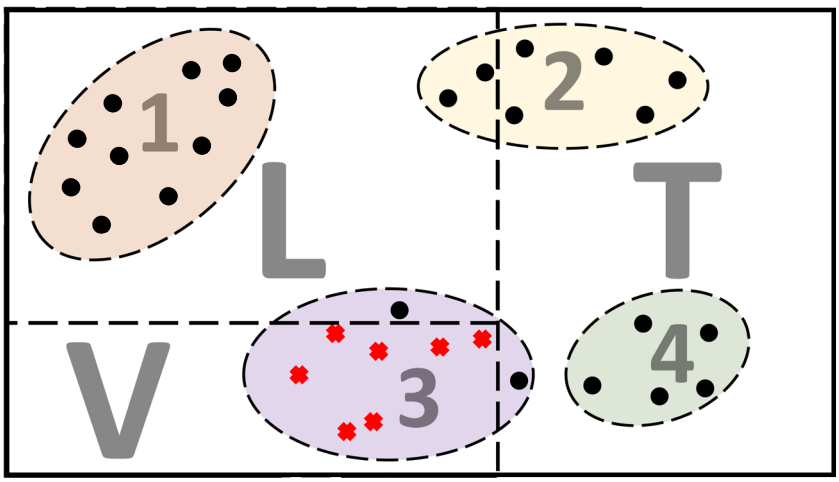

Fig. 3 Sketch of a clustered data set in a double-split situation. Cluster 3 is in validation leakage, while clusters 1, 2 and 4 are not. Red crosses represent validation leakage elements, while black points represent cluster elements that are not. Clusters 2 and 3 are in testing leakage, while clusters 1 and 4 are not. Although other configurations are possible (clusters shared between L and V but not T, or shared between T and V but not L), the figure represents only two kinds of overlap among the four possible for the sake of readability.

that, overall, the distribution of Y knowing $Z = z$ is indeed Mult. Hypg.$(\ell; r_1 - z_1, r_2 - z_2, \ldots, r_{n_c} - z_{n_c})$. □

Corollary 4 *The expected number of validation leakages is*

$$\mathbb{E}[M] = \ell \cdot \left(1 - \frac{1}{n} \sum_{i=1}^{n_c} r_i \frac{\binom{n-r_i}{m}}{\binom{n-1}{m}} \right). \tag{5}$$

Proof As in the single-split case, one has $Y_i = M_i + Y_i \mathbb{1}_{\{Z_i=0\}}$. Summing over clusters and taking expectations as in the proof of Corollary 1, one finds

$$\mathbb{E}[M] = \ell - \sum_{i=1}^{n_c} \mathbb{E}\left[Y_i \mid Z_i = 0\right] \mathbb{P}[Z_i = 0].$$

Using Theorem 2 and our first property of multivariate hypergeometric distributions, one gets $Y_i \mid Z_i = 0 \sim$ Hypg.$(\ell; r_i, n-m)$ and $Z_i \sim$ Hypg.$(m; r_i, n)$. One concludes by expressing and rearranging the summands above. □

From the comparison of Eqs. 3 and 5, let us remark how similar both considered expected proportions of leakage elements appear. Actually, in the single-split situation, $n - k$ is the size of the learning set, while m plays this role for the double-split situation. However, to evaluate the expected number of testing leakages in a double-split situation, we use Eq. 3, in which case the size of the learning set becomes $m + \ell$. In particular, we obtain the following Corollary, which states that, under certain conditions, there is no way to achieve the same expected proportion of testing leakages whatever the size of the validation set. In other words, there is no hope of fully compensating for the impact of leakage induced by clusters by playing on proportions between learning, validation, and test sets.

Corollary 5 (Curse of data leakage) *If there is at least one cluster cardinality satisfying $2 \leq r_i \leq k + \ell$, then the expected proportion of validation leakages is strictly smaller than the expected proportion of testing leakages.*

Proof As V is non-empty, $\ell > 0$, and therefore,

$$\frac{(k-1)!}{(k-r_i)!} \leq \frac{(k+\ell-1)!}{(k+\ell-r_i)!}, \tag{6}$$

for all clusters such that $1 \leq r_i \leq k$, with equality if $r_i = 1$ and strict inequality otherwise. Two cases are distinguished. The first one is when there is at least one cluster cardinality satisfying $2 \leq r_i \leq k$. From Eq. 6, one has

$$\begin{aligned} \sum_{i:\, r_i \leq k} r_i (n-r_i)! \frac{(k-1)!}{(k-r_i)!} &< \sum_{i:\, r_i \leq k} r_i (n-r_i)! \frac{(k+\ell-1)!}{(k+\ell-r_i)!} \\ &\leq \sum_{i:\, r_i \leq k+\ell} r_i (n-r_i)! \frac{(k+\ell-1)!}{(k+\ell-r_i)!}. \end{aligned}$$

For the second case, assume that there is at least one cluster cardinality such that $k < r_i \leq k + \ell$. Then, from Eq. 6,

$$\begin{aligned} \sum_{i:\, r_i \leq k} r_i (n-r_i)! \frac{(k-1)!}{(k-r_i)!} &\leq \sum_{i:\, r_i \leq k} r_i (n-r_i)! \frac{(k+\ell-1)!}{(k+\ell-r_i)!} \\ &< \sum_{i:\, r_i \leq k+\ell} r_i (n-r_i)! \frac{(k+\ell-1)!}{(k+\ell-r_i)!}. \end{aligned}$$

In both cases, this is equivalent to

$$\mathbb{E}\left[\frac{N}{k}\right] = 1 - \frac{1}{n} \sum_{i=1}^{n_c} r_i \frac{\binom{n-r_i}{m+\ell}}{\binom{n-1}{m+\ell}} > 1 - \frac{1}{n} \sum_{i=1}^{n_c} r_i \frac{\binom{n-r_i}{m}}{\binom{n-1}{m}} = \mathbb{E}\left[\frac{M}{\ell}\right],$$

which conclude the proof. □

Remark Equality between the expected proportions of validation and testing leakages is reached when all clusters are either single individuals ($r_i = 1$) or sufficiently large relative to the prescribed validation and testing set sizes ($r_i > k + \ell$).

3 Synthetic experiments

In this section, simulated numerical experiments are performed, and theoretical results of Section 2 are used to interpret the results. After some combinatorial simulations, a regression setup is studied for two different splitting strategies. The sizes and parameters used for datasets generation were chosen to be close to those of the agronomic case study that will be treated in Section 4.

3.1 Combinatorial simulations

Figure 4 shows the graph of the expected number of testing leakages depending on k with results of 1′000 draws of N. Each simulation emulates the random split of a hypothetical data set consisting of $n_c = 174$ clusters of constant size $r = 3$ and computes the number of leakage elements in the test set. The average of the 1′000 simulations plus or minus twice the corresponding standard error are reported in green in Fig. 4, with the corresponding theoretic values based on Corollaries 1 and 2 in black. In the same Figure, the theoretical values based on Corollaries 4 and 6 (see appendix) depending on ℓ are represented via black dashed lines. The latter concern validation leakages after a test split of $k = 104$

(corresponding to roughly 20% of the data set). The average of the 1′000 simulations of M plus or minus twice the corresponding standard error are also reported in red. Note that for each simulation of the validation split, a new test set is also drawn.

Other simulations were conducted in the case of heterogeneous cluster sizes (not shown). In all performed simulations, the results are in complete agreement with the theoretical formulae derived in the previous section.

3.2 Friedman data set with uniform test split

In this subsection, we consider a regression setup where the testing set will be split uniformly among the subsets of k elements of the data set. For a clustered data set, one can think about this setup in at least two predictive situations. The first one is when the predictive goal is to predict a new response for a cluster we already observed, in which case random splitting could be appropriate considering Fig. 2. The second situation is when the aim is to predict a new response for a new (unseen) cluster. In that case, random splitting is not appropriate and the test set is no longer representative given the aim to be achieved. Hence, the following example can be considered an example of both situations, each with a distinct predictive goal.

The data considered here is a clustered version of the synthetic data set described by Friedman (1991). Specifically, $n = 504$ five-dimensional input data are drawn independently from a mixture of Gaussian distribution with $n_c = 168$ component means generated uniformly in the hypercube $[-10, 10]^5$ and covariance matrix $\sigma^2 I$ for all components,

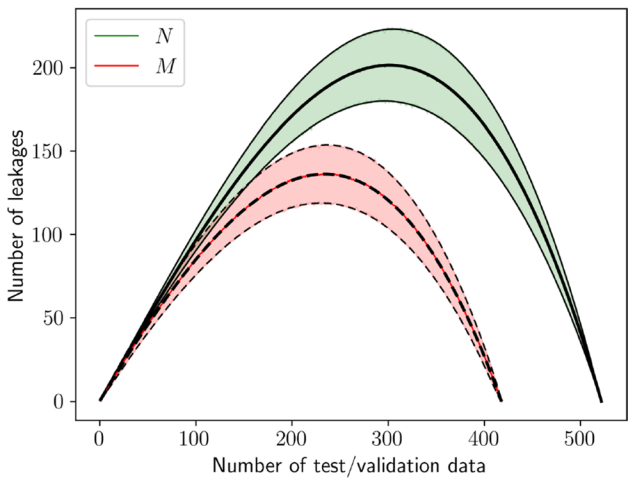

Fig. 4 Number of leakage elements. In black solid line, $\mathbb{E}[N]$ and $\mathbb{E}[N] \pm 2 \cdot \sqrt{\mathrm{Var}[N]}$ for $n = 522$ and constant cluster size $r = 3$. In green, plus/minus twice the standard deviation around the average of 1'000 simulations of N. In black dashed line, $\mathbb{E}[M] \pm 2 \cdot \sqrt{\mathrm{Var}[M]}$ for $k = 104$ surrounded in red by plus/minus twice the standard error around the average of 1'000 simulations of M

with $\sigma = 0.1$. All clusters have a constant size $r = 3$. The input data are then normalized within the hypercube $[0, 1]^5$, and the output data are obtained by the noisy evaluations of the function

$$y(\mathbf{x}) = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5, \qquad (7)$$

where $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ are the input variables, and the noise is independent and Gaussian with mean 0 and variance 1. The top left panel of Fig. 5 displays the first two input variables x_1 and x_2.

A test set of $k = 101$ elements is kept apart by random splitting, corresponding to about 20% of the data set. Then, the KNN algorithm is trained on the remainder of the data. The number of neighbours is selected via a 10 times repeated random validation procedure. The random validation split proportion is varying from about $q = 5\%$ to $q = 50\%$, with 5% steps ($\ell = 21, 41, 61, 81, 101, 121, 142, 162, 182$ and 202). The whole test and validation splitting procedure is repeated 50 times. Averages with two standard errors for the validation and testing MSE are reported in the top right panel of Fig. 5. Firstly, we observe that the test error average is systematically lower than the validation error average. This could be explained by Corollary 5. Indeed, the expected proportion of testing leakages is 96%, while the expected proportion of validation leakages ranges from 94% when $q = 5\%$ to 64% when $q = 50\%$. Secondly, the validation error increases with q. Although this may be due to the reduction of expected leakages when q increases causing the validation error inflation, there may be other reasons for this behaviour (Cawley and Talbot 2010). In particular, we will see in Subsection 3.3 that this behaviour still occurs even when no cluster is in validation leakage.

An analogous experiment is conducted by varying the cluster size from 2 to 9. We set $k = 101$, corresponding to 20% of the whole data set, and $\ell = 81$, corresponding to 20% of the remainder. The $n = 504$ sample is partitioned into a varying number of clusters of different sizes according to Table 2. Here, the data set differs for each 50 repetition and each cluster size. However, the cluster size is constant for each data set, except for the case $n_c = 100$, where 96 clusters have size 5, and 4 clusters have size 6 to reach the same sample size. Results are reported in the bottom left panel of Fig. 5. The average validation error is always higher than the average test error, likely for the same reason as the previous experience. In a consistent manner, and according to expected proportions $\mathbb{E}[N/k]$, $\mathbb{E}[M/\ell]$ provided in Table 2, the cluster size increasing implies a leakage increase, and validation and testing average error curves decreases. Note also that as the cluster size increases, the difference in expected proportion decreases, and the difference between average validation and testing curves tends to decrease.

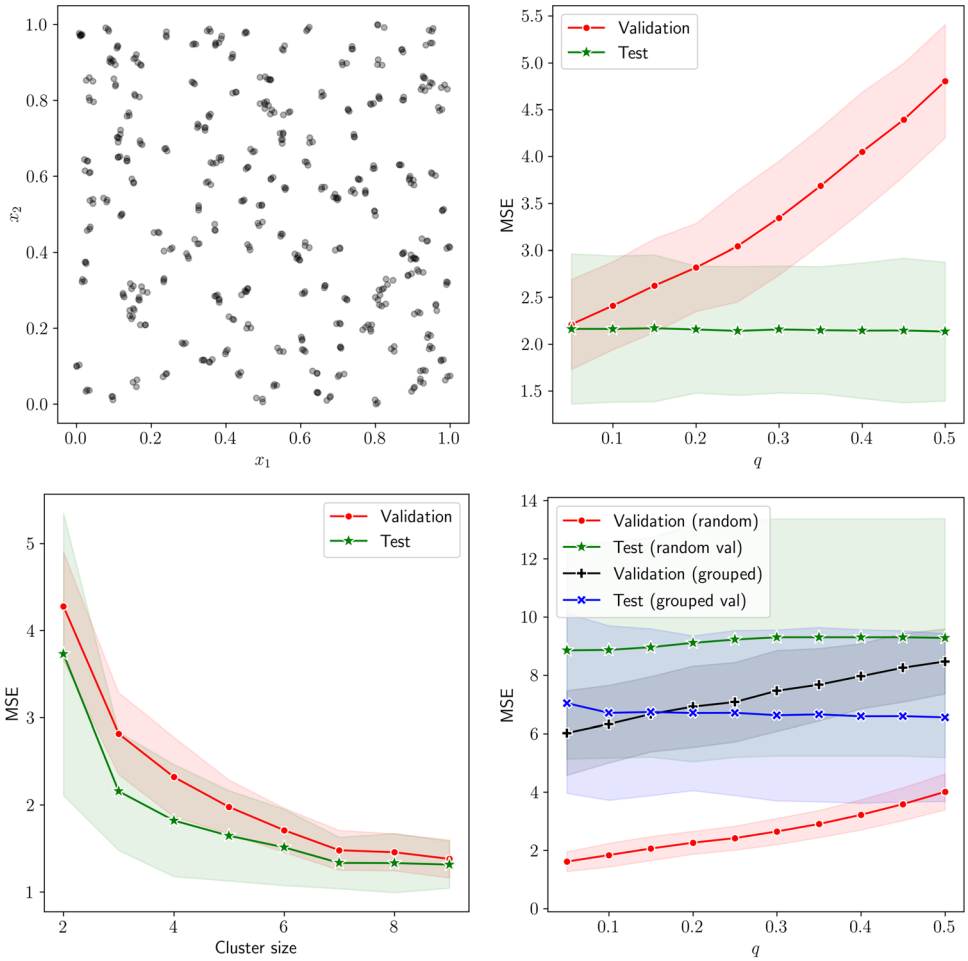

Fig. 5 Experiments with the Friedman data set. (top left) First two input variables of the $n_C = 168$ clusters of size $r = 3$. (top right) Average MSE $\pm$ twice the standard error for the validation and testing set, depending on the validation proportion q, (bottom left) on the cluster size, with constant sample size. (bottom right) Same experiment as in the top right panel with grouped splitting strategy

Table 2 Varying cluster size with the synthetic data set. The data set is drawn with different cluster sizes, corresponding to different expected proportions of testing and validation leakages. The difference of expected proportion is always positive, due to Corollary 5

r	$\mathbb{E}[N/k]$	$\mathbb{E}[M/\ell]$	$\mathbb{E}[N/k] - \mathbb{E}[M/\ell]$	80.0%
2	252	64.0%	16.0%	96.0%
3	168	87.0%	9.0%	99.2%
4	126	95.3%	3.9%	99.8%
5	100	98.3%	1.5%	100.0%
6	84	99.4%	0.6%	100.0%
7	72	99.8%	0.2%	100.0%
8	63	99.9%	0.1%	100.0%
9	56	100.0%	$< 0.1\%$	$k = 102$

3.3 Friedman data set with a grouped test split

A third experiment is done with the same clustered Friedman data set but splitting the test set in a grouped fashion, i.e. no cluster is in testing leakage. Hence the test set will correctly assess the model by estimating its generalization error when the goal is to predict a new response for a new unseen cluster.

More specifically, 34 clusters are kept apart for the test set, corresponding to $k = 102$ elements. The remainder of the data is used for train and validation, according to 10 repeated grouped validation procedures. That is, as with the test split, all validation splits are performed in such a way that each cluster is either in the training or the validation set. A random validation split is also performed for comparison, following the procedure described in Subsection 3.2. The whole experiment is repeated 50 times, and results are shown in the bottom right panel of Fig. 5.

The validation error resulting from the randomly split validation procedure (in red) exhibits behaviour close to that of the experiment when the test was randomly split (see top right panel of the same Figure), which is unsurprising. The models selected from this procedure produced a corresponding test error (in green), which is, on average, 2 to 5 times greater than the validation error. Note that this situation — the grouped testing split with random validation split — is the analogue of the agricultural one presented in the introduction of this paper. One now focuses on the

validation error from the grouped validation procedure (in black). Although the clusters are no longer in validation leakage, the validation error still increases with q, similar to the validation error produced by the randomly split validation procedure. This suggests that this behaviour is not necessarily related to validation leakages, as already mentioned in Subsection 3.2. The corresponding test error (in blue) is better than the testing error resulting from the randomly split validation, indicating better modelling. Although the validation error with the grouped strategy is higher than the one with the random strategy, it is also far more representative of the test error. Using the random strategy for validation results here in an overfitting situation.

4 Application to agronomical data

This section revisits the specific case of the feasibility study of an agronomic recommendation system presented in the introduction. The analyst has to find out whether the system can be developed on the basis of a given data set. After the motivation and presentation of the data set, the data set is uniformly split in order to observe the potential leakage and compared with the theoretical values. We will then follow our analyst through four splitting scenarii in which they will use KNN for the modelling. We will see that the output of the study can vary depending on the strategy used.

4.1 Motivation and data set

For agronomists and plant breeders, the ability to characterize new varieties of crops is important to assess their performance under various conditions. Additionally, the ability to identify where a new variety of crop may be best suited for production based on location characteristics, like weather and soil, is important for maximizing yield, quality, and economic potential of the crop at the farm level. Statistical approaches that identify the relationship between varieties and their environment can help agronomists and growers to find the most appropriate variety for a particular location.

The data set comes from a wheat variety trial network from Agroscope in Switzerland and has been used in Herrera et al. (2018). The network includes 10 locations distributed across Switzerland's main wheat production area. At each location, 6 winter wheat varieties were chosen. The data set contains 3 years of yield measurements (2011, 2012 and 2013). Each combination of year and location defines an *environmental condition*, and each combination of environmental condition and variety is called an *experiment*. Each experiment is done in three replicates. One environmental condition is missing (29 spatio-temporal combinations are available), and one experiment has a missing replicate. Therefore, the data set contains $n = 521$ observations. Additionally, 16 environmental limiting indices specifically designed for winter wheat yield are provided (Herrera et al. 2018; Holzkämper et al. 2013). The 6 variety names are dummy encoded (Murphy 2012). The 22-dimensional input space is composed of variety names and environmental limiting indices. The predictive goal of the study is to predict winter wheat yield for an unobserved environmental condition, given one of the six varieties.

4.2 Data splitting

The nature of the experimental design suggests hierarchical clustered structured data. Therefore, there is a high risk of data leakage during the splitting of the data set. First, experiment replicates are reduced to a singleton for each experiment, which is a particular case of a cluster. Indeed, replicates are the same points for a given experiment in the input space. Second, all experiments conducted within the same environmental conditions could induce potential clusters. This could be a problem, as the goal is predicting yield for an unobserved environmental condition. Incidentally, remark that all experiments conducted with the same variety could also induce potential clusters. However, as the predictive goal here does not include new varieties, this should not be an issue.

The data are uniformly split into three subsets, namely learning, validation and testing set, to select and assess several models in the remainder of the paper. We keep $k = 104$ observations for the testing set (i.e. roughly 20% of the data) and $\ell = 83$ observations for the validation set (i.e. roughly 20% of the remainder of data, $q \approx 20\%$). The number of leakage elements N and M are counted for the validation and testing sets, respectively, and for all potential cluster structures discussed above. The splitting is repeated 10'000 times.

Figure 6 presents the distribution of N and M when the clusters induced by experiment replicates are considered ($n_c = 174$). For N, two behaviours emerge. This is caused by the presence of two distinct cluster sizes. Indeed, since $k \leq n_c$, it is possible to find a split such that the number of testing leakages is k by taking the test set such that all its element belongs to a different cluster. To obtain the nearest smaller realisation of N, all members of one of the smallest clusters must be in the test set. As an example, doing this reasoning recursively for constant cluster size r and $k \leq n_c$, the number of testing leakages can only take the values $k, k-r, k-2r, ...$, as the dominant behaviour in Fig. 6 (left). The secondary behaviour is caused by the cluster of size two and is less noticeable as we have only one cluster of size two against 173 clusters of size three. For the validation set, this multiple behaviour is not perceptible in Fig. 6 (right), as clusters take on various sizes between 1 and r after the first split. The empirical mean and variance of these

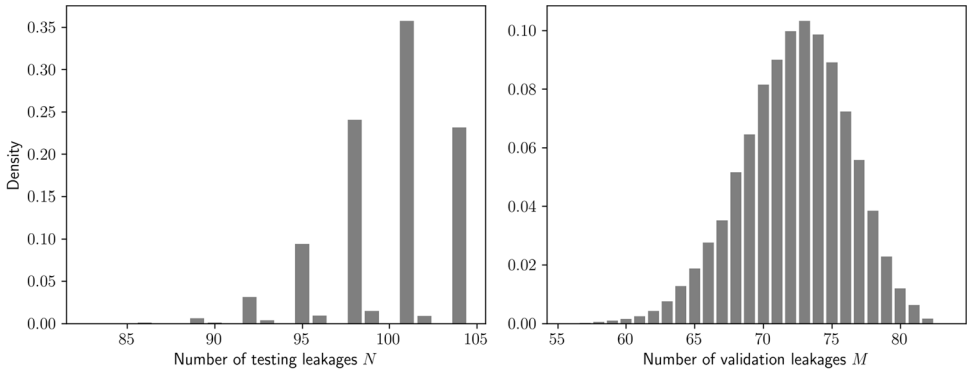

Fig. 6 Simulated distribution of the number of leakage elements induced by experiment replicates. (left) For testing leakages. (right) For validation leakages

Table 3 Mean and variance of the number of leakage elements. Comparison between theoretical values derived in Section 2 and empirical values computed on 10'000 splits of the agronomical data

	Testing set		Validation set	
	mean	var.	mean	var.
Empirical	99.921	10.927	72.287	15.072
Theoretical	99.888	11.014	72.344	15.141

Table 4 Error for the different splitting scenarii of the agronomical data. Mean squared error average (standard deviation) over 50 train/test splits ($n = 174$) and (grouped) 5-fold cross-validation procedures

Scenario	Validation MSE	Test MSE
1	108.03 (14.64)	71.06 (21.19)
2	180.45 (9.32)	195.20 (38.32)
3	70.83 (12.71)	525.88 (161.67)
4	251.36 (27.28)	270.80 (100.58)

simulated distributions are in agreement with the theoretical values derived in Section 2, as reported in Table 3. Moreover, remark that by applying Corollary 3, one concludes that there is a very high probability of having leakage, which is confirmed by the simulated distributions.

The results are not shown graphically for the other potential cluster structure induced by environmental conditions, as the number of elements in the testing leakages (validation leakages) almost always equal k (respectively ℓ). This is in agreement with the fact that $\mathbb{E}[N/k]$ (respectively $\mathbb{E}[M/\ell]$) is very close to 100%.

4.3 Modelling with KNN

We now consider the regression of the wheat yield on the explanatory variables. The KNN algorithm with Euclidean distance is chosen to illustrate our purpose.

4.3.1 Modelling scenario 1 - The analyst erroneously draws uniformly the testing and validation sets

A test set of $k = 104$ elements is drawn uniformly among the subsets of k elements of the data set. The number of neighbours is chosen with a classical 5−fold cross-validation procedure where the random splitting is done uniformly, which corresponds to $\ell = 83$ for three folds and $\ell = 84$ for two folds ($q \approx 20\%$). The procedure is repeated 50 times. The number of neighbours selected is always 1, and the validation and test errors are shown in Table 4. The test error average is lower than the validation error average. This is in accordance with what we observed in Subsection 3.2, and the fact that $\mathbb{E}[N/k] \approx 96\% > 87\% \approx \mathbb{E}[M/\ell]$.

The analyst believes that they are achieving good results on the test set by splitting the data randomly and comparing results to the yield variance (286.85). However, the model has been tested/validated on replicated of experiments that have already been seen during training. The impact of this mistake is evaluated in modelling scenario 3. However, this is only a mistake with respect to the predictive goal. Indeed, the random split of the data corresponds to the case where the analyst's goal is to predict yield for a variety and environment already known to the algorithm, although this hypothetical situation is not of practical agronomical interest.

4.3.2 Modelling scenario 2 - The analyst erroneously draws the testing and validation sets by grouping the replicates

Assume that the analyst is aware that one should not validate/test the model on replicates and makes sure to keep measurements from the same experiment grouped together during all splittings. This procedure ensures that each cluster induced by replicates belongs to one of the split subset (train, validation or test) and is not in leakage. Among the 174 experiments, 35 are used for the test set, and the number of neighbours is selected with a 5-fold group cross-validation. Over 50 repetitions of this procedure, the selected

number of neighbours ranged from 37 to 49 neighbours. Results are shown in Table 4.

Unlike modelling scenario 1, the test error average is higher than the validation error. At first sight, this could be interpreted as a sign that there is no more leakage. Let us provide another explanation in the light of the theoretical results in Section 2. Under the grouping strategy at the replicate level, the formulae (3) and (4) can be reused to count the approximate expected number of experiments in the testing/validation set that has at least one experiment with the same environmental conditions by setting $n = 174$, $r = 6$, $k = 35$, and $\ell = 28$ for four folds, $\ell = 27$ for the remaining fold. We get $\mathbb{E}[M/\ell] \approx 99\%$ and $\mathbb{E}[N/k] \approx 100\%$, indicating that most of the time, all groups of replicates in the validation and testing sets have at least a group in the training set that have the same environmental conditions. That is, information might still be leaked via potential clusters induced by environmental conditions, but $\mathbb{E}[N/k] - \mathbb{E}[M/\ell]$ is too small to allow the validation error to be higher than the test error.

4.3.3 Modelling scenario 3 - The analyst splits the test data accordingly to the predictive goal but erroneously draws the validation set uniformly

This situation corresponds to simulations of Subsection 3.3 and to the introductory example. Here, the test set is properly split to assess the model according to the predictive goal by grouping the data with the same environmental conditions so that the environmental conditions of the test set have not been already seen during the training/validation phases. However, the analyst selects the number of neighbours, always equal to 1, with a classical 5-fold cross-validation procedure. We can also think of this situation as a way of seeing the impact of data leakage if the model is developed with uniformly random splitting by assuming that the test set is unknown to the analyst. According to Table 4, the analyst receives good results during the model development (validation error) and might believe the model will do a good job, while the model will be completely off once in production (test error is almost 2 times the yield variance). This can have disastrous practical consequences.

Incidentally, we remark that the validation error is lower than for modelling scenario 1, while the same multiple-fold cross-validation strategy is used. That might be explained by the fact that, considering the clusters induced by replicates, modelling scenario 1 is a double split situation requiring formula (5) which gives $\mathbb{E}[M/\ell] \approx 87\%$ of leakages while modelling scenario 3 is a single split situation requiring formula (3) yielding $\mathbb{E}[N/k] \approx 96\%$. This indicates that there are more validation leakages due to replicates in modelling scenario 3, which may artificially improve the validation error.

4.3.4 Modelling scenario 4 - The analyst split the test data accordingly to the predictive goal and correctly do the corresponding validation procedure

Finally, we consider the case of the analyst aware of all the pitfalls mentioned so far, avoiding any leakage situation. The data with the same environmental conditions are grouped for the testing split and the 5-fold cross-validation procedure. Therefore, environmental conditions in the test set have not been already seen during the training/validation phases and the validation procedures mimic this situation. Thus, the number of neighbours, ranging from 15 to 216, is properly selected and the model is assessed in accordance with the predictive goal. This corresponds to simulations of Subsection 3.3. The results are reported in Table 4 after 50 repetitions. The validation error is representative of the test error. Considering the test error variance and comparing the test error with the yield variance, the analyst can very likely conclude that there is not enough structure in this dataset to achieve the predictive goal. This is a very different conclusion from the modelling scenario 1.

5 Conclusion

This paper discusses how and to which extent clustered data are allocated after data splitting. Based on this allocation, the validation or/and testing set may be easier to predict at, which potentially yield overoptimistic results depending on the predictive goal. This can have an impact on both model selection and model assessment. The leakage induced by clusters between the different subsets has been described by probabilistic modelling. Under the assumption of uniform drawing of the subsets, analytical results have been derived supported by numerical simulations and empirical findings. While these derivations were done in the context of a single validation or test set, formulae for the expected number of leakages are still valid for multiple-fold cross-validation procedures.

The present agronomic case study demonstrates the impact of cluster-induced data leakage in the presence of inadequate splitting. In this actual data set, clusters are induced by the experimental design. As a consequence, a naive splitting procedure makes it easier for the model to predict the validation and test data, misleading the hyper-parameters optimisation process and the evaluation of the model's predictive performance. Depending on the splitting strategy, the analyst moved from a virtually ideal situation (scenarii 1 and 2), to a situation where the initial objective should be abandoned if no further data are available (scenario 4).

Although this paper clarifies the mechanisms of data leakage in the presence of clusters, this is undoubtedly a simplified view of the problem. Predicting the outcome of a leaking situation is challenging and relies on the specific

models and data involved. Investigations on model comparisons under data leakage could be the subject of future research.

In conclusion, when the data contains a known cluster structure, it is essential to leverage this information to ensure the reliability of model selection and evaluation. On the other hand, in cases with an unknown or ignored cluster structure, the analytical results reveal a high proportion of leakage elements in most situations. This yields a higher risk of misleading generalization error estimation and inadequate hyperparameter tuning.

Appendix A Variance of the number of leakages in the double-split situation

$$\mathrm{Var}[M] = \frac{\ell(\ell-1)}{n(n-m-1)}\left(\sum_{i=1}^{n_c} r_i(r_i-1)\frac{\binom{n-r_i}{m}}{\binom{n-1}{m}} + \sum_{i\neq j} r_i r_j \frac{\binom{n-r_i-r_j}{m}}{\binom{n-1}{m}}\right) + \frac{\ell}{n}\sum_{i=1}^{n_c} r_i \frac{\binom{n-r_i}{m}}{\binom{n-1}{m}} - \left(\frac{\ell}{n}\sum_{i=1}^{n_c} r_i \frac{\binom{n-r_i}{m}}{\binom{n-1}{m}}\right)^2.$$

Corollary 6

Proof Mimicking proof of Corollary 2, one finds

$$\begin{aligned}\mathrm{Var}[M] &= \sum_{i=1}^{n_c}\sum_{j=1}^{n_c}\mathrm{Cov}\left[Y_i\,\mathbb{1}_{\{Z_i=0\}}, Y_j\,\mathbb{1}_{\{Z_j=0\}}\right]\\ &= \sum_{i=1}^{n_c}\mathbb{E}[Y_i^2 \mid Z_i=0]\,\mathbb{P}[Z_i=0]\\ &\quad + \sum_{i\neq j}^{n_c}\mathbb{E}[Y_i\,Y_j \mid Z_i=0,\,Z_j=0]\,\mathbb{P}[Z_i=0,\,Z_j=0]\\ &\quad - \left(\sum_{i=1}^{n_c}\mathbb{E}\left[Y_i \mid Z_i=0\right]\mathbb{P}[Z_i=0]\right)^2.\end{aligned}$$

We know that $Z_i \sim \mathrm{Hypg.}(m;r_i,n)$, $Y_i \mid Z_i=0 \sim \mathrm{Hypg.}(\ell;r_i,n-m)$, $(Z_i,Z_j,m-Z_i-Z_j)^T \sim \mathrm{Mult.Hypg.}(m;r_i,r_j,n-r_i-r_j)$, and $(Y_i,Y_j,\ell-Y_i-Y_j)^T \mid (Z_i,Z_j,m-Z_i-Z_j)^T=(0,0,m) \sim \mathrm{Hypg.}(\ell;r_i,r_j,n-m-r_i-r_j)$, by Theorem 2 and the third of our list of properties on the multivariate hypergeometric distribution. One obtains the end result via the first and second properties from the latter list and some elementary algebra. □

Acknowledgements The authors are grateful to Dr. Amanda Burton, Dr. Didier Martial Pellet, and Dr. Riccardo Turin for the profitable discussions. They would also like to thank the reviewers for their constructive comments, which helped to improve the paper.

Author contributions F.G. and D.G. conceived the main conceptual ideas, developed the theoretical formalism, discussed the results, and wrote the original draft. The computational results were performed and interpreted by F.G. The supervision was carried out by D.G. In addition, L.L.H. and J.M.H. designed and run the field trials, collected, and curated the data. All authors, provided critical feedback, commented, reviewed, and edited the original manuscript, corrected the final version of the paper, and gave final approval for publication.

Funding Open access funding provided by University of Bern. This work was supported by Agroscope, swiss granum, the Swiss Federal Office for Agriculture, the University of Bern, the Schweizerischer Getreideproduzentenverband, Prometerre, Jowa SA and Timac Agro Swiss.

Data availability The agricultural datasets analysed during the current study are available from Dr. Lilia Levy Häner (lilia.levy@agroscope.admin.ch) on reasonable request.

Declarations

Competing interests The authors declare no competing interests.

References

Ayotte B (2021) Fast user authentication via keystroke dynamics (Unpublished doctoral dissertation). Clarkson University

Ayotte B, Banavar MK, Hou D, Schuckers S (2021) Group leakage overestimates performance: a case study in keystroke dynamics. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1410–1417

Buntaran H, Piepho H-P, Hagman J, Forkman J (2019) A cross-validation of statistical models for zoned-based prediction in cultivar testing. Crop Sci 59(4):1544–1553. https://doi.org/10.2135/cropsci2018.10.0642

Cawley GC, Talbot NL (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 11:2079–2107

Friedman JH (1991) Multivariate adaptive regression splines. Ann Statist 19(1):1–67. https://doi.org/10.1214/aos/1176347963

Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, vol 2. Springer

Herrera JM, Levy Häner L, Holzkämper A, Pellet D (2018) Evaluation of ridge regression for country-wide prediction of genotype-specific grain yields of wheat. Agric For Meteorol 252:1–9

Holzkämper A, Calanca P, Fuhrer J (2013) Identifying climatic limitations to grain maize yield potentials using a suitability evaluation approach. Agric For Meteorol 168:149–159. https://doi.org/10.1016/j.agrformet.2012.09.004

Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions, vol 165. Wiley New York

Kapoor S, Narayanan A (2023) Leakage and the reproducibility crisis in machinelearning-based science. Patterns 4(9):100804. https://doi.org/10.1016/j.patter.2023.100804

Kaufman S, Rosset S, Perlich C, Stitelman O (2012) Leakage in data mining: formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data (TKDD) 6(4):1–21

Kuhn M, Johnson K (2019) Feature engineering and selection: a practical approach for predictive models. CRC Press

Meghnoudj H, Robu B, Alamir M (2023) Sparse dynamical features generation, application to parkinson's disease diagnosis. Eng Appl Artif Intell 126:106882. https://doi.org/10.1016/j.engappai.2023.106882

Montesinos López OA, Montesinos López A, Crossa J (2022) Multivariate statistical machine learning methods for genomic prediction. Springer Nature

Murphy KP (2012) Machine learning: a probabilistic perspective. MIT press

Nisbet R, Elder J, Miner GD (2009) Handbook of statistical analysis and data mining applications. Academic press

Rabinowicz A, Rosset S (2020) Cross-validation for correlated data. J Am Stat Assoc 117(538):718–731. https://doi.org/10.1080/01621459.2020.1801451

Rice JA, Silverman BW (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. J Roy Stat Soc: Ser B (Methodol) 53(1):233–243. https://doi.org/10.1111/j.2517-6161.1991.tb01821.x

Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8):913–929. https://doi.org/10.1111/ecog.02881

Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Ser B Methodol 36(2):111–147