Precision Agriculture
https://doi.org/10.1007/s11119-019-09650-0

Automatic harmonization of heterogeneous agronomic and environmental spatial data

Corentin Leroux[1,2] · Hazaël Jones[2,3] · Léo Pichon[2] · James Taylor[2] · Bruno Tisseyre[2]

Abstract
The analysis and mapping of agronomic and environmental spatial data require observations to be comparable. Heterogeneous spatial datasets are those for which the observations of different datasets cannot be directly compared because they have not been collected under the same set of acquisition conditions, for instance within the same time period (if the variable of interest varies across time), with consistent sensors or under similar management practices (if the management practices impact the measured value) among others. When heterogeneous acquisition conditions take place, there is a need for harmonization procedures to make possible the comparison of such observations. This analysis details and compares four automated methodologies that could be used to harmonize heterogeneous spatial agricultural datasets so that the data can be analysed and mapped conjointly. The theory and derivation of each approach, including a novel, local spatial approach is given. These methods aim to minimize the occurrence of discrepancies (discontinuities) in the data. The four approaches were evaluated and compared with a sensitivity analysis on simulated datasets with known characteristics. Results showed that none of the four methods consistently delivered a better harmonization accuracy. The accuracy and preferred choice for the harmonization procedures was shown to be influenced by (i) within-field spatial structures of the datasets, (ii) differences in acquisition conditions between the heterogeneous spatial datasets, and (iii) the spatial resolution of the simulated data. The four approaches were used to harmonize real within-field grain yield datasets and a discussion to help users select an appropriate harmonization methodology proposed. Despite significant improvements in dataset harmonization, discontinuities were not entirely removed and some uncertainty remained.

Keywords Data harmonization · Discontinuity · Spatial autocorrelation · Yield

✉ Corentin Leroux
cleroux@smag.tech

Bruno Tisseyre
bruno.tisseyre@supagro.fr

[1] SMAG, Montpellier, France

[2] ITAP, Univ Montpellier, Irstea, Montpellier SupAgro, Montpellier, France

[3] MISTEA, Univ Montpellier, INRA, Montpellier SupAgro, Montpellier, France

Published online: 20 March 2019

Introduction

Large amounts of data are being acquired within fields using precision agriculture technologies. From satellite platforms to embedded or manual field sensors, data are collected with the intent of helping agricultural professionals and producers to characterize within-field spatial variability and to make informed management decisions (Oliver 2010). Usually these data are processed as a whole within a relatively small spatial extent (e.g., at the field scale), using commonly reported spatial methods of data analysis, as it is assumed that these data were collected under homogeneous acquisition conditions. This assumption is often wrongly made, for various reasons (discussed later). Acquisition conditions that are necessary for the integrity of this assumption might be classified into five major groups:

- *Time-related* data collected in a relatively short amount of time to limit the influence that temporal variations in the sensing environment can have on the acquired data (e.g., temperature variations between the beginning and end of acquisition process),
- *Sensor-related* data collected with multiple sensors, but with similar calibration settings to minimize measurement biases,
- *Operator-related* data collected by multiple operators, but with common operating procedures, to minimize operator-dependent bias (handheld sensor, potential effect of machine handling on the measurement for embedded sensor),
- *Method-related* data collected with similar measurement methods or underlying models to ensure observations represent the same information,
- *Management practices-related*: data collected under similar management practices, (training systems, sowing date) so that the important external spatial factors affecting the crop growth are not confounded by management effects.

If the integrity of data, as qualified by these five conditions, cannot be guaranteed, then observations cannot be directly compared as these observations belong to heterogeneous datasets (Baume et al. 2009; Brenning et al. 2008; Fassó et al. 2007). Some typical examples include:

- Soil apparent electrical conductivity measured in neighbouring fields with different underlying soil moisture conditions (e.g., fallow vs cropped conditions) that influences the sensor response. The result is problematical for mapping soil properties across field boundaries (Weller et al. 2007),
- To expedite harvesting, multiple combines are used within the same field, each of them using yield monitors with different calibration settings (Maldaner et al. 2016; Sams et al. 2017),
- Mapping of water status over many fields of large farms using vegetation indices as surrogate measures of soil water status in Mediterranean conditions (Acevedo-Opazo et al. 2008). However, management practices also impact on vegetation indices (e.g., plantation date, variety, training systems, cover-cropping) and need to be accounted for when merging fields under varied management regimes.

To compile and analyse spatial heterogeneous information, processing techniques are required to reconcile differences represented by differing acquisition conditions/methods. These techniques are referred to as harmonization procedures. They essentially aim at

correcting and rescaling data so that heterogeneous datasets can be merged and/or compared (Baume et al. 2009; Köhl et al. 2000). The process is not the same as standardization or data fusion. Standardization is a primary procedure such that all the methods and sources to collect observations have been made comparable. It is conceivable to assume that standardization cannot be always achieved which is why complementary approaches, such as harmonization, are required (Baume et al. 2009). Data fusion generally refers to methods that combine data collected at different resolutions (e.g., spectral, spatial or temporal), to generate more accurate and reliable information. In this sense, harmonization processes are as a subset of data fusion methods.

Harmonization procedures are needed in many spatial application domains, such as soil, vegetation, or health sciences (Bartholomeus et al. 2008; Brenning et al. 2008; Fassó et al. 2007). Corrective methodologies have been proposed to harmonize data before further analyses. Simple but still efficient algorithms have tended to compare either global or local statistics of neighbouring heterogeneous datasets to generate weighted corrective factors that were used to harmonize several adjacent datasets (Maldaner et al. 2016; Weller et al. 2007). Sams et al. (2017) used referenced observations to harmonize spatial data arising from different harvesters with different calibration settings. Even though the use of reference data could be considered as an optimal solution, such reference data are not always available. Other authors have come up with spatial approaches, mostly using kriging, to compare adjacent data and contrast with heterogeneous data (Baume et al. 2010; Brenning et al. 2008; Skøien et al. 2010). Some of these methods made use of a harmonization function that was selected to minimize the difference between an interpolated value from a reference dataset and a harmonized value from an adjacent heterogeneous dataset (Brenning et al. 2008). Others required a proper modelling of the spatial structure of the data to reconstruct the spatial autocorrelation (Baume et al. 2010). However, these last methods are hardly automatable, especially because they require a cautious fit of a theoretical variogram to the data.

As sensing systems and data collection increases in agricultural systems, the issue of data heterogeneity will become a more important issue that potentially limits the utility of these new agricultural datasets. The major purpose of this analysis is therefore to detail and contrast four automated (unsupervised) approaches to help harmonize heterogeneous spatial agricultural datasets, particular for cases where the discontinuity in the data is due to the differences in acquisition conditions. This includes the comparison of two existing approaches and the proposition of two novel approaches. A key emphasis is on automated (unsupervised) methods, as these will be required in the near future to process the predicted increase in the number of heterogeneous datasets. As data increases, the ability to harmonise data manually will decrease.

Merging heterogeneous spatial datasets collected under known acquisition conditions

Harmonization of heterogeneous spatial datasets: theoretical aspects

Consider a regionalized agronomic variable Z defined over a domain D that a user would like to map over this same domain D. Within the domain, $Z(x)$ can be considered as a random variable and a realization of the variable Z at the specific location x. At all locations x within D, this variable can be collected by a sensor (e.g., satellite, unmanned aerial

vehicle, mounted sensor, human, etc.), either directly or indirectly. It might be possible that the exact same information as Z is sensed, which means that Z is obtained instantaneously (e.g., operator directly measures the leaf area index [LAI] by taking hemispherical images of canopy). Otherwise, an alternative regionalized variable Y, which is somehow related to Z, is sensed because measurements of Z are constrained by time and/or cost considerations (e.g., a satellite-based vegetation index that is used to estimate LAI). In the latter case, which happens often in practical situations, an agronomic model has to be set so that the values of Z can be derived from the values of Y. This agronomic model can be formalized by a transformation f that relates the observations of Y and Z (Eq. 1):

$$\hat{Z}(x) = f(Y(x)) \tag{1}$$

where $\hat{Z}(x)$ is an estimate of the variable Z at the position x.

Note that this equation also stands for the case in which Z is sensed directly (i.e., f would be the identical function, assuming the sensor is accurate with comparable precision). Now, consider that the regionalized variable Y has been sensed under two different acquisition conditions over the domain D (Fig. 1). These two heterogeneous datasets Y_1 and Y_2 might have been collected for instance with different machines, sensors, operators or at different times. These spatial datasets can be categorized into two groups, spatially separated or nested in space (Fig. 1). In the first case, observations are collected in different portions of the field, (e.g., soil apparent electrical conductivity measurements collected at two different dates). In this case, because different conditions existed during data acquisition, a clear discontinuity is observed between the datasets (Fig. 1a).

In the second case, observations belonging to different datasets are mixed in space (e.g., crowdsourcing observations collected by two different operators working in close proximity to each other in a field). Here, the discontinuity is unconstrained and appears at many places across the field (Fig. 1b). As these datasets were collected over the same field, it is conceivable to assume that the available observations should contain relatively consistent values. Yet, agricultural spatial datasets generally exhibit some sort of spatial

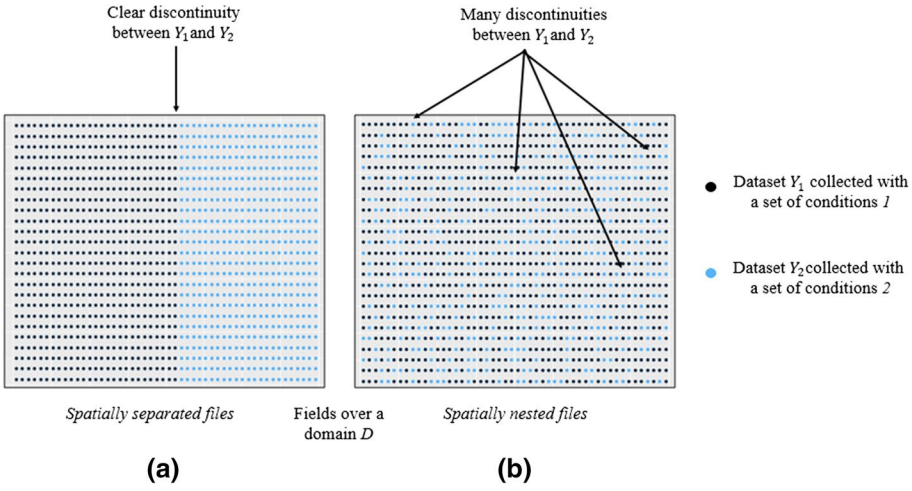

Fig. 1 Two spatial datasets collected under different acquisition conditions. Datasets can be **a** spatially separated (left) (e.g., soil electrical conductivity data in neighbouring fields) or **b** nested in space (right) (e.g., crowdsourced data)

autocorrelation, to a greater or lesser extent, meaning that the variations of Z (and Y) would be expected to be somewhat continuous over the field. However, the discontinuity as illustrated raises questions as to whether these two sets of observations belong to the same population. This discontinuity is important to consider and to take into account because it is likely to mask the inherent patterns of variation within the field if not corrected. In this analysis, focus was given to datasets spatially separated (the condition indicated in Fig. 1a).

Because each set of acquisition conditions implies a specific representation of Y, the measurements of Y in these two datasets cannot be compared. The regionalized variable Y_i is denoted to represent the variable Y under the specific set of acquisitions conditions c_i. The relation between Y_i and Y can be written as follows:

$$\hat{Y}(x) = g_i(Y_i(x)) \tag{2}$$

where $\hat{Y}(x)$ is an estimate of the variable Y at the position x, and g_i is the transformation function that relates the variables Y_i to Y.

A theoretical example of two datasets with differing responses for the same variable that require harmonization is given in Fig. 2. Y_1 and Y_2 differ in both the mean and variance of their response to measurements of Y. To harmonize these data, transformation functions g_i are needed. These functions can have multiple forms. However, in many cases, these relationships can be summarized by linear functions. In fact, these approximations through linear modelling are interesting because they suggest that the set of acquisition conditions c_i implies a shift in the mean value of the Y_i and a shift in the variance of the Y_i (Fig. 2). The first shift can be understood as the bias of the sensor (a global and systematic offset for each measurement which can be seen as the linear model intercept). The second shift can be viewed as the sensitivity of the sensor (the smallest absolute amount of change that can be detected by a measurement which can be seen as the linear model slope). Although

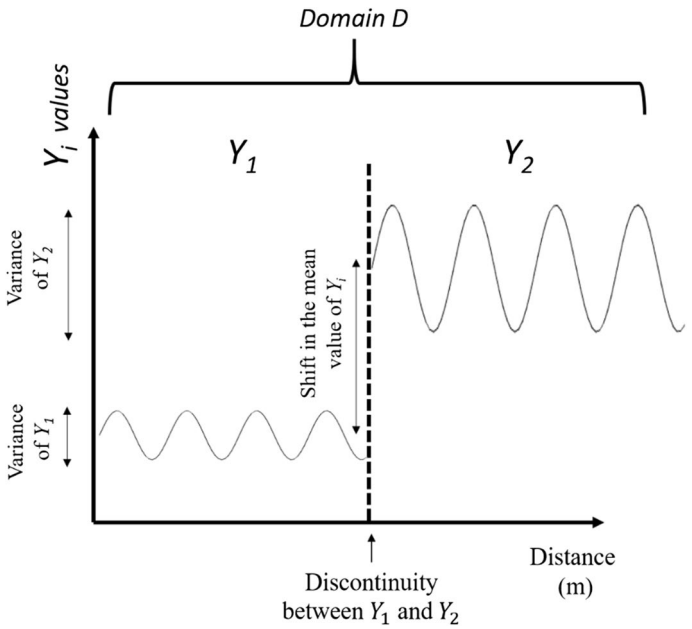

Fig. 2 Theoretical shifts in the mean and variance of measurements of Y_i because of differences in acquisition conditions (c_1 and c_2) that give rise to two different, non-harmonious populations (Y_1 and Y_2)

more complex functions can be set, linear approaches are still appropriate because they can embrace a significant part of the relationship between Y_i and Y (Eq. 3).

$$\hat{Y}(x) = a_i \ * \ Y_i(x) + b_i + \varepsilon \tag{3}$$

where a_i and b_i stand respectively for the slope and intercept of the linear function g_i, ε is the error term and corresponds to the accuracy of the sensors.

If all the transformations g_i are known, estimates of Y can be retrieved by simply using the functions g_i. If the g_i functions are not known, all the Y_i variables should be harmonized with respect to the same reference. The reference may be one of the Y_i chosen randomly or the selection of one Y_i that is known to be more accurate, such as when one sensor calibration is known to have been done properly and uncertainty exists for the other sensors. In any case, the application of g_i obtains an estimate of Y and there would then be a need to calibrate with a specific *f* function to retrieve the values of Z (Eq. 1)

So far, it has been considered that g_i were stable and applied over the domain *D*. However, nothing prevents the parameters of g_i from evolving in space or in time. To be more specific, it would be possible to define Eq. 3 as:

$$\hat{Y}(x,t) = a_i(x,t) \ * \ Y_i(x,t) + b_i(x,t) + \varepsilon \tag{4}$$

where x and t stand for the spatial location and temporal acquisition date of a given observation, ε is the error term.

For simplification purposes, this latter case was not addressed with this analysis, and the parameters of the linear function were considered stationary in space and time. This implied from a variogram perspective that the range of autocorrelation was considered stable over space (Fig. 3).

Four approaches to harmonize heterogeneous spatial datasets

Here four methods to harmonize spatial datasets are described. These were chosen for being easily computable and automated. The first two approaches have been used previously (Maldaner et al. 2016). The third methodology is proposed as a complement to the

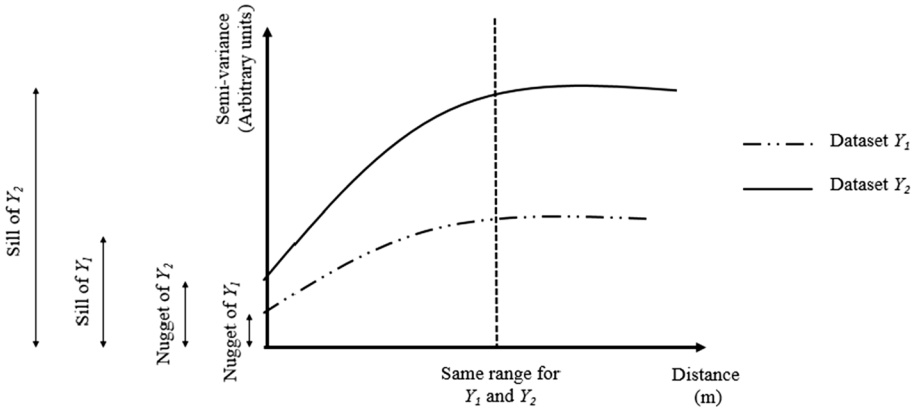

Fig. 3 Impact of the differences in acquisition conditions of the Y_i on their variogram representation. Note that the range was considered constant to avoid non-stationary considerations

first two approaches. Contrary to the first three, the last approach is a novel method that accounts for the spatial relationships in the data in the harmonization process.

Consider two neighbouring heterogeneous datasets Y_1 and Y_2 that represent the variable Y ($g_1(Y_1)$ and $g_2(Y_2)$), over a given spatial extent (Fig. 4). Consider that Y_1 is selected as the reference dataset. In this case, g_1 can be considered as the identical function. Therefore, the objective would be to determine the parameters of the linear function g_2 in order to harmonize Y_2 with respect to Y_1.

A simple global methodology: M_{Glob}

This first approach simply aimed at centring the values of Y_2 with respect to Y_1 (Maldaner et al. 2016). Here, the linear function g_2 only contains a slope parameter, a_2, defined as:

$$a_2 = \frac{\bar{Y}_2}{\bar{Y}_1} \tag{5}$$

where $\bar{Y}_1$ and $\bar{Y}_2$ are the mean values of the whole datasets Y_1 and Y_2 (Fig. 4).

A simple local methodology: M_{Loc1}

The second approach was very similar to the previous methodology except that the parameter a_2 was computed within a local neighbourhood near the discontinuity between the heterogeneous spatial datasets (Fig. 4). Even though this approach was relatively simple, it considered that the data near the discontinuity should be more related than when the entire datasets were taken into account (Maldaner et al. 2016). The size of the local neighbourhood used is somewhat arbitrary. Ideally it should be restricted to an area (distance) within which autocorrelation is known to occur. Some even more local methodologies, such as point-to-point comparisons under the condition that the local neighbourhood near

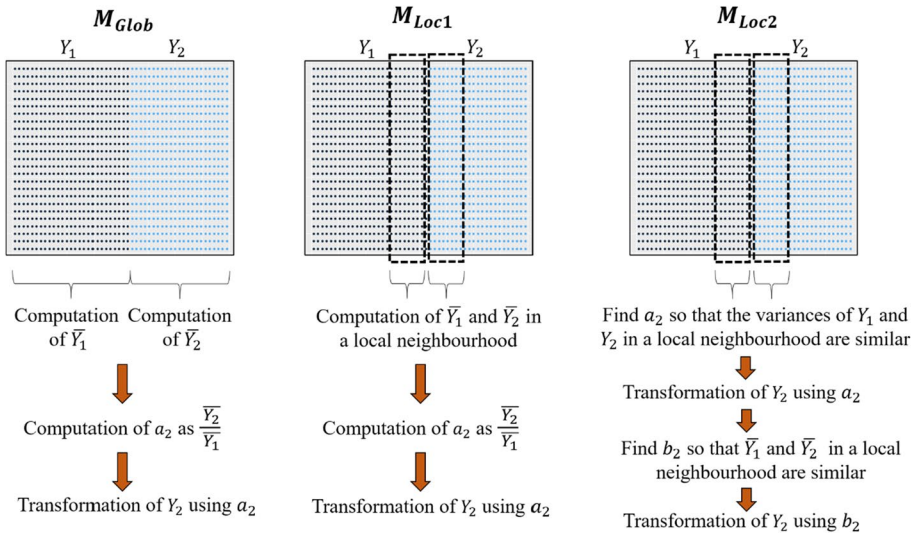

Fig. 4 Simple non-spatial methodologies to harmonize heterogeneous datasets

the discontinuity is restricted to a single point, have been proposed (Maldaner et al. 2016; Weller et al. 2007), but this special condition was not considered here.

An advanced (novel) two-step local methodology: M_{Loc2}

A third novel methodology was proposed as the two previous approaches do not account for possible differences in variance between datasets Y_1 and Y_2. The distribution of both datasets was expected to be similar, but it is possible that the acquisition conditions generated a narrower or wider distribution (as illustrated in Fig. 2). First, this approach scaled Y_2 with respect to Y_1. The objective was to find the parameter a_2 of the linear function g_2 through an optimization procedure so that the variances of Y_2 and Y_1 were aligned. Then, a second step aimed at centring the resulting scaled Y_2 dataset with respect to Y_1 to harmonize the mean of both datasets (Fig. 4). The objective was to find the parameter b_2 of the linear function g_2 so that both mean values of Y_2 and Y_1 were similar.

A simplified spatial methodology: M_{SP}

The proposed simplified spatial approach M_{SP} intends to account for the spatial relationships that exist between adjacent or nested heterogeneous spatial datasets (Y_1 and Y_2). This method was said to be "simplified" because it was not based on proper modelling of spatial structure, especially because automation is the main target.

The hypothesis is that a consideration of the within-field spatial structure will help to improve the harmonization accuracy. As differences in acquisition conditions should not affect the spatial autocorrelation, the difference in attribute value between two observations of Y_1 separated by a spatial distance d_i should be similar to the difference in attribute value between an observation of Y_1 and an observation of Y_2, separated by the same spatial distance d_i (Eq. 6). The proposed method aimed to find the parameters of a minimization function $m(a_2, b_2)$ so that the spatial autocorrelation across Y_1 and Y_2 is maintained (Eq. 6). The calculation was made for distances (d) lower than the range of autocorrelation of Y_1 because there should not be any correlation between two observations once this distance is reached. Additionally, as autocorrelation between observations was expected to be stronger for observations nearer in space than for those further away, a weighting scheme w_d was proposed. As d increased, the weight associated with the correction was lowered (Eq. 6)

$$m(a_2, b_2) = \sum_{d=0}^{range} w_d * \frac{1}{2N(d)} \sum \left| \left| g_1(Y_1(x) - g_1(Y_1(x+d)) \right| - \left| g_1(Y_1(h) - g_2(Y_2(h+d)) \right| \right| \tag{6}$$

where $N(d)$ are the number of pair of points separated by a distance d, w_d is a weighting scheme, and x and h are spatial locations in space.

Given that Y_1 was selected as the reference dataset, and that g_2 was a linear function, the minimization function $m(a_2, b_2)$ can be rewritten as follows:

$$m(a_2, b_2) = \sum_{d=0}^{range} w_d * \frac{1}{2N(d)} \sum \left| \left| Y_1(x) - Y_1(x+d) \right| - \left| Y_1(h) - (a_2 Y_2(h+d) + b_2) \right| \right| \tag{7}$$

In Eq. 6, no restriction was made on the distribution of the Y_2 values. Indeed, the function $m(a_2, b_2)$ could be minimized either with a narrow or wide distribution of the Y_2 values. This might be problematical as the distribution (variance) of Y_1 and Y_2 is usually expected to

be similar. To cope with this issue, Y_2 was scaled with respect to Y_1 following the first step proposed in M_{Loc2}. The minimization function $m(a_2, b_2)$ of Eq. 6 was then simply $m(b_2)$ as the parameter a_2 was known with the previous scaling step. It must be added that the decision to first scale Y_2 with respect to Y_1 was also made because the function $m(a_2, b_2)$ exhibited many local minima in initial testing and it was difficult to propose an automatic robust harmonization procedure to find the global minimum of both parameters simultaneously.

The automation of this approach necessitates the automatic determination of the variogram range. Automated variogram fitting can be problematic and this is a potential limitation to the M_{SP} approach. However, the exact range value is not theoretically needed, as the weight w_d associated with large distances will be very low, and the principal function of the range parameter is to restrict the analysis to a relevant neighbourhood. The range of 'average' variograms (the range found on a variogram from a typical field or several fields in the area) or existing known variogram ranges could be substituted instead (McBratney and Pringle 1999). Other authors have recently proposed advanced spatial harmonization algorithms mostly relying on the comparison between the kriged estimates from one dataset and the true values of the second dataset to harmonize (Baume et al. 2010; Skøien et al. 2010). However, even though these last methods are valuable, it was decided to exclude these for this analysis as they require user parametrization and supervision for proper modelling of spatial structure.

Materials and methods

Simulated spatial datasets

When multiple methods are proposed to cope with harmonization of data, a difficulty exists with how to evaluate if the approaches are able to provide an accurate data correction. In fact, without reference observations, uncertainty persists about whether the different acquisition conditions were correctly considered. One way of tackling this issue is to use simulated datasets with known properties and behaviour and to mimic varying acquisition conditions (Brenning et al. 2008; Leroux et al. 2017). Given that the initial simulation conditions are known, it is much easier to evaluate how effective a proposed methodology would be in reconciling the issues associated with the conditions. Simulated datasets can thus be used to compare multiple harmonization approaches, and therefore direct guidance for best options with real datasets.

Spatial datasets were created using the R statistical environment (R Core Team, Vienna, Austria). The objective was to evaluate the quality of the methodology on controlled and known datasets first, before applying it to real agronomic spatial datasets. These simulated datasets were 4-ha square fields (200 * 200 m) with a mean of approximately 7 (arbitrary units). Spatially correlated datasets were simulated via the sequential simulation algorithm in the *gstat* package (Bivand et al. 2013). The coefficient of variation was set to 40% and the nugget to sill ratio of the variogram was set to 40% as these values can be found in generic agronomic or environmental datasets (Pringle et al. 2003). Theoretical variograms were modelled with spherical functions. The other descriptive aspatial and spatial statistics used in the simulations (i.e., data resolution, range of the variogram, and sensor sensitivity) are detailed in the sensitivity analysis. These initial simulated datasets represented the variable Y. Simulated datasets were then divided into two equal subsets to account for differences in acquisition conditions. Observations in one of these subsets were linearly transformed using a linear function

g_2^{-1}. The linearly transformed datasets were denoted Y_2 and the other subset, which was left unchanged, was denoted Y_1. Note that the transformation here is g_2^{-1} and not g_2 because g_2 is the linear function to harmonise Y_2 with respect to Y_1. As indicated previously, the two datasets were only considered to be separated in space along a clear line of discontinuity (Fig. 1a).

Parametrization and evaluation of the non-spatial and spatial approaches

For the local approaches M_{Loc1} and M_{Loc2}, the local neighbourhood was defined as the observations lying within a distance less than the variogram range from the discontinuity. For the simulations, the range of the variogram was known as it was an input variable used to generate the simulated datasets.

The local approach M_{Loc2} and the simplified spatial approach M_{SP} required an optimization procedure to find the parameters a_2 and b_2. The determination of these parameters was done in R using a one-dimensional optimization approach implemented in the function "*optimize"* following the work of Brent (1973) in the R package "stats". This approach was selected for being widely used in one dimensional optimization problems. Regarding the proposed spatial approach, an inverse distance weighting was set for the weighting scheme w_d (Eq. 8).

$$w_d = \frac{1}{d^p} \tag{8}$$

where p, the power parameter, sets the influence of observations separated in space by a distance d. The higher the values of p, the stronger the influence of nearer observations in space. In this study, p was set to 2 to provide a substantially greater influence of observations nearer in space in the minimization function $m(a_2, b_2)$ but still allowing observations further apart to contribute to the harmonization procedure. The determination of w_d is arbitrary, and could equally be estimated from the variogram, although this is more difficult to automate effectively. For the practicalities associated with automation, the inverse distance approach was preferred here.

Regarding the simulated datasets. the initial true values of Y were known as the whole dataset was simulated with specific aspatial and spatial characteristics (Table 1). Therefore, the quality of the correction was assessed by computing the average error of estimation of the Y values (Eq. 9).

$$Error = 100 * \frac{mean(|\hat{g}_2(Y_2) - Y|)}{\bar{Y}} \tag{9}$$

where $\hat{g}_2$ is the estimate of g_2, $\bar{Y}$ is the mean of Y.

Table 1 Input parameters in the sensitivity analysis for the harmonization of heterogeneous spatial datasets

Type	Criterion	Definition	Associated values
Spatial structure	Range	Maximum distance of autocorrelation	40 m (small range) 120 m (high range)
Sensor characteristics	a_2 (slope of g_2^{-1})	Measurement sensitivity	1.5
	b_2 (intercept of g_2^{-1})	Measurement offset	1 (small bias) 6 (high bias)
Data resolution	Data resolution	Number of points per hectare	100 (small resolution) 1000 (high resolution)

This error term was calculated for the observations of Y_2 lying within a distance less than the range of the variogram to the observations of Y_1, and a condition imposed on all four methods for harmonizing the results. The objective was to evaluate the error term near the discontinuity that existed between the datasets Y_1 and Y_2.

All the methodologies were further evaluated by testing them on simulated datasets with variations in (i) the within-field spatial structure represented by the variogram range, (ii) the acquisition conditions (i.e., bias and sensor sensitivity modelled by the parameters a_2 and b_2 of the linear transformation g_2^{-1}) and (iii) the spatial resolution (density) of the data. Values used in the sensitivity analysis are reported in Table 1. For each non-spatial and spatial characteristic considered, 50 datasets were simulated. The outcomes of the four harmonization procedures on the simulated datasets was evaluated through a box-plot analysis. All the criteria were tested independently one at a time.

Real-world dataset

Harmonization methodologies were also applied to a real-world yield dataset obtained from grain flow sensors mounted on combine harvesters. Yield data were considered a very good case-study for the proposed methodologies as these are likely to be split into separate spatially discrete datasets. For instance in large fields, harvest may be performed by multiple combine harvesters operating at the same time. In cases where the calibration of the sensors inside these harvesters is different, there is a need to provide a correction so that all the data can be read and plotted at the same time. It is also possible that the harvesting is done over several days by a single machine. Here again, a correction may be needed to make sure that the combined dataset exhibits the true patterns of yield variations within the field. This second case study was considered here. The data used was from a 7.5 ha canola field in 2004 located near Alnwick, England. The field was harvested using a Claas combine (Harsewinkel, Germany) with a swath width of 5 m over a two-day interval. Before applying any harmonization procedure, the yield dataset was filtered to remove outliers and inliers in the data (Leroux et al. 2018).

Results and discussion

Characteristics of spatial datasets collected under different acquisition conditions

Simulated data mostly followed a gaussian distribution before the function g_2^{-1} was applied to the right-hand portion of the field (Fig. 5, left). In this example, the spatial structure was characterized by a nugget to sill ratio of 60% and a range of approximately 60 m. When the linear transformation g_2^{-1} was applied to the right portion of the field, both spatial and non-spatial distributions of the data were affected (Fig. 5, right). The global distribution effectively started to lose its gaussian shape for a more skewed distribution. Note that, with real data, this skewed distribution could be due to spatially varying environmental phenomena, such as distinct and differing soil parent material. Agri-environmental data does not have to be normally distributed. However, in this case the acquisition conditions were known to be different between the two datasets. With larger values of slope and intercept in the function g_2^{-1}, the distribution would have been more affected and ultimately bimodal in nature. The within-field spatial structure was also substantially affected by the g_2^{-1} transformation as (i) the nugget and variance increased within the field, and (ii) the stationarity of the process

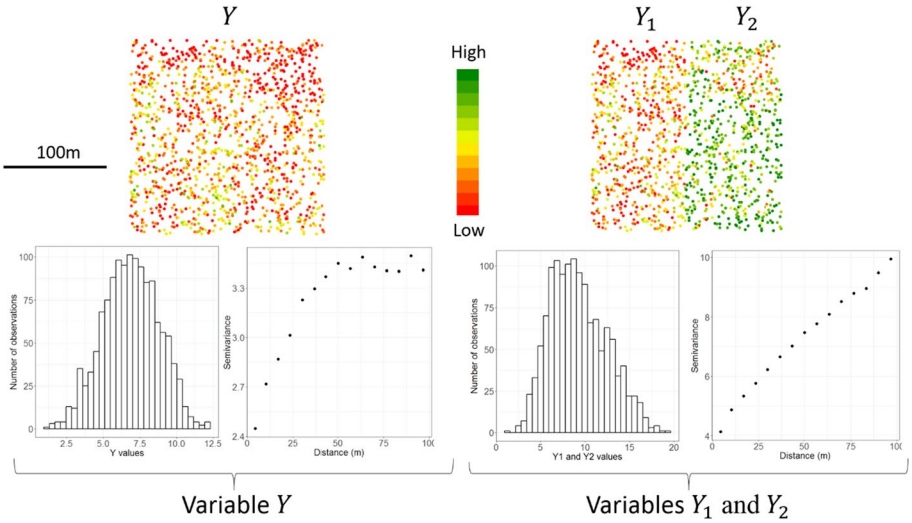

Fig. 5 Simulated dataset before (left) and after (right) a linear transformation g_2 was applied to the right portion of the field. A histogram and experimental variogram of the data are shown to illustrate the difference in the classic statistical and geostatistical nature of the two datasets. Of note, histograms are on different x-axes and variograms are on different y-axes for visualisation purposes

disappeared. Figure 5 demonstrates how important it is to determine accurately the parameters of the function g_2 to remove the influence of different acquisition conditions on the spatial and aspatial distribution of the data.

Sensitivity analysis of the harmonization procedures

Figure 6 reports the accuracy of the harmonization procedure for the four methods for varying (i) within-field spatial structures (range) (Fig. 6a, b), (ii) differences in sensor operation (bias) (Fig. 6c, d), and (iii) spatial resolution of the data (Fig. 6e, f). Since these results show relative difference and have no statistical test per se, interpretation of error rate differences should be tempered. This limitation is raised because error was controlled with the simulated datasets. For instance, if the coefficient of variation in the simulated datasets was set to a higher value, the harmonization errors would also increase significantly (data not shown). Results show that no single approach was preferred. The harmonization accuracy of the four methodologies was dependent on the inputs and conditions imposed on the simulated datasets. Similar conclusions have been obtained when comparing several harmonization procedures on real within-field yield datasets (Maldaner et al. 2016). This simulation analysis was interesting in the sense that even simple automatic harmonization approaches in some cases performed better than more complex spatial methods. In other words, implementing a more complex automatic harmonization method that accounts for data autocorrelation does not guarantee better accuracy.

All three varying parameters had an influence on the accuracy of each harmonisation method (Fig. 6). Regarding the within-field spatial structure, low-spatially structured fields (i.e., small variogram range) tended to favour non-spatial harmonization methodologies. Indeed, the simplified spatial method, M_{SP}, generated a higher median and a

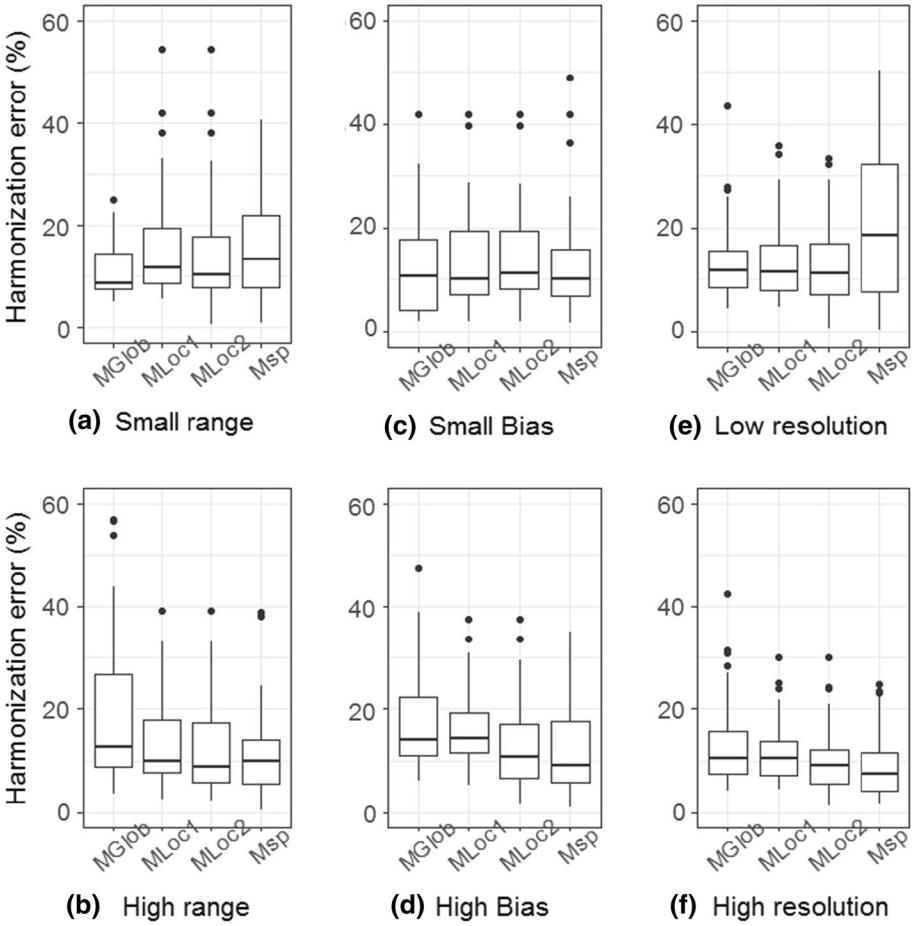

Fig. 6 Impact of the within-field spatial structures (range) (**a**, **b**), sensor performance (bias) (**c**, **d**), and spatial resolution of the data (**e**, **f**) on the harmonization error for the four methods under investigation [Fifty simulations were run for each varying parameter]. The midline is the median of the data. Whiskers extend up to 1.5 times the interquartile range from the top (bottom) of the box to the furthest datum within that distance. Data (points) beyond that distance represent outliers

wider distribution of the absolute error of correction compared to the non-spatial methods. The simple global method M_{Glob} was found the most reliable in this case as the use of a global corrective factor tended to smooth local variations. Results were completely reversed for well-spatially structured datasets (Fig. 6). This finding was not surprising given that the relationship between spatial observations decreased when the distance between these observations increased, which was taken into account in M_{SP}. Note that the distribution of the M_{SP} errors in this case was also much narrower. In the case of known low-spatially structured datasets, one solution to improve the results of the M_{SP} approach would be to decrease the power parameter (p) of the weighting scheme (w_d) so that very close spatial observations do not have an overwhelming influence on the correction.

Figure 6 also demonstrates the impact that the bias of the sensor (linear model intercept) can have on M_{Glob} and M_{Loc1}. As this bias increased, the difference in harmonization accuracy between these two simple approaches and the last two methodologies became much more distinct. As the bias increased, the M_{Glob} *and* M_{Loc1} approaches tended to generate a higher scaling factor a_2 (Fig. 4). By doing so, the distribution of the resulting Y_2 dataset contracted, which increased the harmonization errors. This is an issue of concern if the difference in acquisition conditions between two heterogeneous spatial datasets is not known in advance. If these differences were small, then both M_{Glob} *and* M_{Loc1} outputs could be as reliable as those of the more complex methodologies, such as M_{Loc2}. However, if it happens that these differences are large, then using the first two simple methods might be inadequate. Lastly, regarding the data resolution, it only seemed to impact the simplified spatial method M_{SP} (Fig. 6). Low-spatial resolution data should not be processed with M_{SP}. When few observations are available, the spatial structure cannot be accurately reconstructed by the M_{SP} approach, which leads to a higher harmonization error. This last approach is not recommended for datasets where the spatial structure is based on a low number of observations (Webster and Oliver 1992).

While harmonization methods helped remove the discontinuities between heterogeneous datasets, these results indicate that it was impossible to create perfectly harmonized datasets. Indeed, some errors and uncertainties still remained inside the harmonized datasets and it is important to account for this artifact when mapping or analyzing data (Brenning et al. 2008). From a general perspective, there is always some level of noise in the data and it is not practical to suggest any methodology to remove this inherent level of error. Figure 6 also shows that relatively high harmonization errors (points outside the boxplots) can be obtained. This might be for instance the case when the data were not continuous over the datasets to be harmonized (i.e., there was a real discontinuity between the datasets), but this aspect might be considered relatively rare in reality. In such cases, all the corrective procedures would lead to a poor correction, but the resulting discrepancy would be obvious in the harmonized map. Some artefacts might also have been generated during the simulation of the datasets (e.g., large differences in either the mean or variance in the data near the discontinuity), which could have led to large harmonization error.

With respect to the underlying question to which this study intended to answer, the conclusion is not obvious given the results obtained. Given the impact of the sensor bias on the accuracy of both M_{Glob} *and* M_{Loc1} approaches, these methods are not recommended, especially if it is not known in advance the difference in sensor bias between the heterogeneous spatial datasets. The proposed spatial method should be used for harmonizing agronomic or environmental datasets known to exhibit substantial spatial autocorrelation. This suggestion only holds if the amount or the spatial density of the data is sufficient for accurate geostatistical analysis, particularly the determination of variogram structures (a simple rule of thumb could be to consider the limit conventionally accepted for the calculation of a semivariogram, $n > 50$) From a more general perspective, if there is absolutely no prior information regarding the dataset to be harmonized, using the advanced local M_{Loc2} approach is recommended. This method delivered the most stable results. From a more practical and operational perspective, the differences between the median error of each method were not particularly large (up to 5%) and it is unclear how much effect this error would have on the resulting spatial patterns of harmonized data. Results will likely vary from an obvious visual effect to being negligible. However, as previously discussed, some harmonization methods did deliver more stable results than others.

Table 2 Yield descriptive (t ha^{-1}) and spatial statistics of the initial and harmonized dataset (right section of the field)

Type	Method	Non-reference set (right portion of the field)				Whole field		
		Min	Mean	Max	Variance	Skewness	C_0	C_1
Initial	–	3.73	4.51	5.19	0.06	0.50	0.04	Non stationary
Harmonized	M_{Glob}	3.05	3.69	4.25	0.04	− 0.22	0.024	0.022
	ML_{oc1}	3.00	3.63	4.18	0.04	− 0.18	0.025	0.022
	M_{Loc2}	2.95	3.63	4.23	0.05	− 0.19	0.027	0.022
	M_{SP}	3.03	3.72	4.31	0.05	− 0.22	0.026	0.022

'Initial' indicates the dataset after technical yield errors have been removed. 'Harmonized' indicates the dataset that has been corrected for different acquisition conditions. C_0 and C_1 are respectively the nugget and partial sill of the variogram model that was fitted to the data

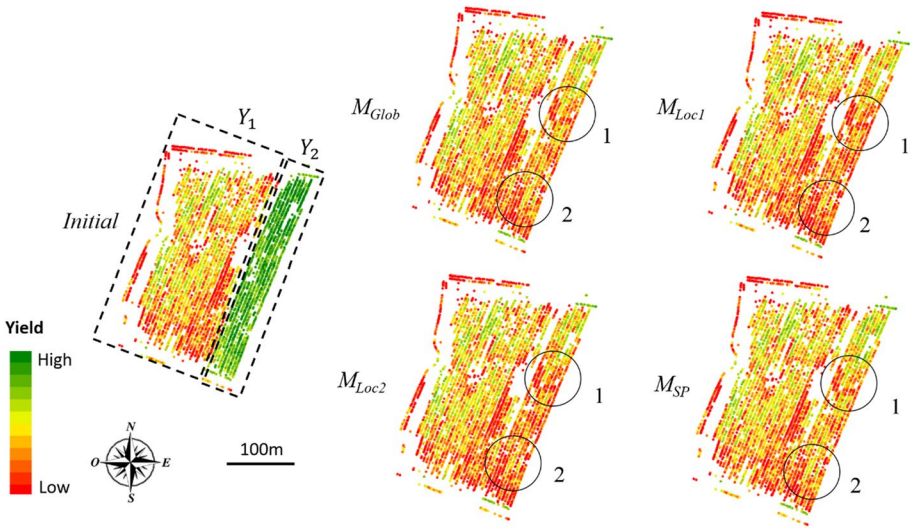

Fig. 7 Evolution of the canola yield spatial patterns before and after the files have been processed. Circles 1 and 2 are areas of interest for which descriptive statistics can be found in Table 3. All data are plotted on a common legend (t ha^{-1})

Applicability of the harmonisation approaches to a real-world yield dataset

Table 2 reports the descriptive and spatial statistics of the real yield dataset before and after the file has been harmonized with the four previously described harmonization methods. All these initial and harmonized yield datasets are plotted in Fig. 7. Visual inspection shows that observations within this field are unlikely to originate from the same population (Fig. 7). A clear discontinuity between the data collected at different dates is noted and labelled as Y1 (western larger portion) and Y2 (eastern smaller portion). Harmonization procedures substantially changed the data distribution in the Y2 (non-reference) part of the field (Table 2). Data skewness after harmonization was much closer to zero, indicating

the data distribution became more gaussian. The four harmonization approaches generated relatively similar global yield descriptive outputs. Larger differences are expected for local patterns/statistics.

The application of the four methodologies to harmonize heterogeneous datasets helped retrieve a global within-field structure as the conditions of stationarity were met after the corrective procedures (Table 2). Variograms effectively became stationary with a nugget to sill ratio between 40 and 50%. Figure 7 also shows that yield patterns appeared to be much more continuous and much smoother across the field. The low yielding area in the southern-portion of the field was reconstructed.

Even though the global yield statistics were found to be similar across the four harmonisation methods, some local divergences could be observed as shown within the delineated circles (Fig. 7). The non-spatial local approaches M_{Loc1} and M_{Loc2} produced lower yield values than the two other methods, certainly because most of the eastern section of Y_1 exhibited a lower yield than the western part of the field (Table 3). Following this reasoning, the global approach produced higher yield values because the high yielding area in the western part of the field was taken into account in the harmonization procedure. When examining the eastern section of Y_1, more especially the eastern two rows of Y1, yield values were slightly higher than the surrounding low yielding area. Given that the M_{SP} approach attributed a higher weight in the harmonization to small distances between observations, this method resulted in relatively higher yield values in the Y_2 dataset. This case study illustrates how the M_{SP} approach was sensitive to the yield observations located directly adjacent to the discontinuity between the heterogeneous datasets, and therefore how pre-processing filtering methodologies need to be accurate. If outliers remain or some sharp discontinuity exists, the quality of the spatial harmonisation procedure can be affected (Weller et al. 2007). Of note, yield variance in both circles was higher for methods M_{Loc2} and M_{SP} (Table 3), that accounted for differences in variance between the left and right-hand parts of the field as discussed in the previous sections.

From a general perspective, it is relatively difficult to estimate the error that might remain after harmonization procedures are applied, and consequently difficult to identify which method produced the most accurate results. However, the proposed corrections have retrieved a probably lost spatial structure and have certainly helped to compute more reliable yield representation and associated spatial statistics for the field. In the case of spatially separated datasets, the accuracy of the harmonization procedure will depend on the number of observations lying near the discontinuity (Weller et al. 2007). For instance, there was one relatively large area in the centre of the field that was void of observations. Though adjacent to the Y1/Y2 boundary, this was not problematical given many other observations were available for the correction. However, if observations are sparse near the discontinuity between spatially-separated data, simple global methods, such as M_{Glob}, would likely be the better harmonization option.

Table 3 Within-field descriptive statistics of yield (t ha^{-1}) in two specific portions of the field after the harmonisation procedures were applied

Circle	Initial		M_{Glob}		ML_{oc1}		M_{Loc2}		M_{SP}	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
1	4.49	0.072	3.68	0.034	3.63	0.033	3.62	0.039	3.71	0.039
2	4.23	0.167	3.62	0.037	3.57	0.037	3.56	0.043	3.64	0.042

How to retrieve the values of the variable Z of interest?

The goal of the harmonization procedure was only to provide an estimate of the variable Y (i.e., $\hat{Y}$; Eq. 2). Recall that the variable Y was introduced because observations of the variable of interest, Z, may not be directly available (e.g., soil organic carbon content (Z) estimated through the collection of soil spectra). As an indirect measurement of Z is made, referred to as Y, a question arises on how to define the parameters of the function *f* that relates Z to Y (as shown in Eq. 1) so that the values of the variable of interest Z can be retrieved?

Situations may exist where the interest is not in the absolute values of Z but in the relative values Z across a field. For instance, relative Z values could be sufficient for delineating within-field management zones. In this case, the function *f* does not need to be determined because the variable Y already contains the required information. However, there are more complex situations in which absolute values of Z are needed. For instance, agronomic yield models need real crop data (e.g., vegetation parameters, not just relative vigor indices) to calibrate and or validate yield predictions. To transform the Y variable, there is a need to have access to some reference values of Z, either at a global or a local scale:

- Global. At a global scale, some descriptive statistics of Z are typically known (e.g., mean, variance) but very few or no values of Z are known for each observation of Y. For instance, the average yield within a field (mean of Z) can be known by weighing the truck(s) that collected the total grain removed from the field. However, the grain (Z value) has not been weighed within the field each time the yield monitor collected a measurement (Y value). In this case, the function *f* can solely be based on the comparison of the descriptive statistics of Y and Z. Regarding the previously described example, it might be possible to compare the mean value of the observations collected by the yield monitor (mean Y) to the average yield arising from the weighing of the truck containing the grain (mean Z).
- Local. At the local scale (e.g., areas within fields) there may exist some collocated measurements of Y and Z. As an example, a model that relates soil spectra (Y) to soil organic carbon content (Z) by measuring in the laboratory soil organic content values of some of the samples for which soil spectra were collected. This model can then be used directly within the field to derive the values of Z when new soil spectra are collected. In this case of collocated measurements, users should make sure that the measurements of Y and Z are acquired within the same spatial support to ensure values of Y and Z can be compared.

Another point of consideration is the selection of a reference value/dataset and is illustrated with the yield map case study. The Y_i chosen as the reference dataset was the one with the highest number of observations. This might also be the one for which the Y values are the most different from those of Z, but as no other information is known regarding the quality of the sensor's calibrations, a choice had to be made. To retrieve the absolute values of the variable Z of interest, (i.e., the real yield within the field) the authors advocate using the total weight of the grain as measured when removed from the field. With this absolute information, users will be able to approximate the function *f* by relating the mean values of Y and the average true yield in the field (this would be a way to account for the sensor's bias of the reference dataset that was chosen). However,

the function f will not be retrieved entirely. The weight of the grain is precisely measured, such that the mean yield is known, however the yield variance within the field remains unknown.

Last considerations

A quality of the proposed simplified spatial methodology is that the spatial distance between neighbouring observations matters. The observed difference between close observations in space will be given more importance in the harmonization procedure than observations far apart. Indeed, as the spatial distance between observations can vary strongly, a need exists to account for this information. The proposed simplified spatial method does not require the data to be fit to a geostatistical model. It is acknowledged that the range parameter needs to be set but the exact range value is not theoretically needed, as the weight w_d associated with large distances will have minimal effect, and the principal function of the range parameter is to restrict the analysis to a relevant neighbourhood. This means that (i) raw data can be used directly and (ii) the harmonization procedure can be automated. In the proposed approach, a weighting scheme w_d (inverse distance power of 2) was put into place to give a major influence to observations nearer in space. Such parametrization generated a relatively small effective neighbourhood, but there was also a need to consider that there were many more pairs of points separated by larger distances. In other words, pairs that represent greater separation distances were given a lower weight but were much more numerous than pairs of near observations. If data exhibits a relatively strong within-field short-scale spatial structure, the use of a local neighbourhood is appropriate and will generate an accurate correction. Problems arise when the data are noisy and the within-field spatial structure is more complex, such as when spatial data that has multiple sill variances. In such case, the inverse distance's power should be lowered so that the local neighbourhood does not have an overwhelming influence on the harmonization procedure. One potentially interesting improvement of the proposed methodology would be to use a weighting scheme that is related to the variogram range of the data to prevent the M_{SP} from being too local. However, in order for this to be automated, a reasonable estimator of the range is needed.

In the simulated and real case studies, only two heterogeneous datasets were considered at one time. Conceivably more than two datasets might need to be harmonized and merged. As such, the correction errors would increase at each iteration, as some uncertainty would propagate with each new harmonization procedure. Also the linear transformation function g_i was considered stationary in space and time (Eq. 3). In this analysis with only two heterogeneous spatial datasets separated in space, the spatial distance was short and the assumptions were considered acceptable. However, if multiple heterogeneous datasets were to be considered together, this assumption of stationarity may need to be challenged. Maldaner et al. (2016) ended up with the same conclusion as the corrective procedures they proposed had spatially varying efficiencies in different areas of the field in their investigation.

Conclusion

Four automated approaches were detailed to harmonize heterogeneous spatial datasets so that observations inside these datasets can be directly compared. Among these, three were considered non-spatial as they did not account for spatial autocorrelation in the data. The

fourth method considered spatial relationships in the data to minimize the occurrence of discrepancies, but was not based on the formal modelling of the spatial structure, as automation was the main target. The use of two simulated datasets demonstrated that none of these algorithms outperformed the others under situations of (i) varying within-field spatial structures of the datasets, (ii) differences in sensors performance between the heterogeneous spatial datasets, and (iii) spatial resolutions of the simulated data. Nonetheless, when working with medium to high-resolution spatial agronomic information, the proposed simplified spatial method was able to provide a more accurate harmonisation correction. A proposed non-spatial local approach did account for possible differences in variance between the heterogeneous datasets, giving the most stable results across all simulations. Even if the discontinuity between neighbouring heterogeneous datasets was substantially removed, some uncertainties remained and should be accounted for when analysing and mapping conjointly heterogeneous datasets. Furthermore, the transformation functions to harmonize the observations were considered stationary in space and time. This aspect would require more investigation if these conditions were not true, especially for very large spatial datasets where conditions of stationarity are by nature often compromised.

Acknowledgements This work, referred as ANR-16-CONV-0004, was supported by the French National Research Agency under the "Investments for the Future Program."

References

Acevedo-Opazo, C., Tisseyre, B., Guillaume, S., & Ojeda, H. (2008). The potential of high spatial resolution information to define within-vineyard zones related to vine water status. *Precision Agriculture, 9,* 285–302.

Bartholomeus, R. P., Witte, J. P. M., van Bodegom, P. M., & Aerts, R. (2008). The need of data harmonization to derive robust empirical relationships between soil conditions and vegetation. *Journal of Vegetation Science, 19,* 799–808.

Baume, O., Skøien, J., Carré, F., Heuvelink, G., & Pebesma, E. (2009). Data harmonization of environmental variables: From simple to general solutions. In J. Hřebíček, J. Hradec, E. Pelikán, O. Mírovský, W. Pilmmann, I. Holoubek, & T. Bandholz (Eds.), *European conference of the Czech presidency of the council of the European Union towards environment* (pp. 162–169).

Baume, O., Skøien, J. O., Heuvelink, G. B. M., Pebesma, E. J., & Melles, S. J. (2010). A geostatistical approach to data harmonization—Application to radioactivity exposure data. *International Journal of Applied Earth Observation and Geoinformation, 13,* 409–419.

Bivand, R. S., Pebesma, E. J., & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R*. New York, NY, USA: Springer.

Brenning, A., Koszinski, S., & Sommer, M. (2008). Geostatistical homogenization of soil conductivity across field boundaries. *Geoderma, 143*(3), 254–260.

Brent, R. (1973). *Algorithms for minimization without derivatives, Chap. 4*. Englewood Cliffs, NJ, USA: Prentice-Hall.

Fassó, A., Cameletti, M., & Nicolis, O. (2007). Air quality monitoring using heterogeneous networks. *Environmetrics, 18,* 245–264.

Köhl, M., Traub, B., & Païvinen, R. (2000). Harmonization and standardization in multi-national environmental statistics—Mission impossible? *Environmental Monitoring and Assessment, 63,* 361–380.

Leroux, C., Jones, H., Clenet, A., Dreux, B., Becu, M., & Tisseyre, B. (2017). Simulating yield datasets: An opportunity to improve data filtering algorithms. In J. A. Taylor, D. Cammarano, A. Prashar, & A. Hamilton (Eds.), *Proceedings of the 11th European conference on precision agriculture*. *Advances in Animal Biosciences, 8,* 600–606.

Leroux, C., Jones, H., Clenet, A., & Tisseyre, B. (2018). A general method to filter out defective spatial observations from yield mapping datasets. *Precision Agriculture, 19,* 789–808.

Maldaner, L. F., Molin, J. P., & Canata, T. F. (2016). Processing yield data from two or more combines. In *Proceedings of the 13th international conference on precision agriculture.* Retrieved March, 2019, from https://www.ispag.org/proceedings/?action=abstract&id=1965&search=years.

McBratney, A. B., & Pringle, M. J. (1999). Estimating average and proportional variograms of soil properties and their potential use in precision agriculture. *Precision Agriculture, 1,* 125–152.

Oliver, M. A. (2010). *Geostatistical applications for precision agriculture*. London, UK: Springer.

Pringle, M. J., McBratney, A. B., Whelan, B. M., & Taylor, J. A. (2003). A preliminary approach to assessing the opportunity for site-specific crop management in a field, using a yield monitor. *Agricultural Systems, 76,* 273–292.

Sams, B., Litchfield, C., Sanchez, L., & Dokoozlian, N. (2017). Two methods for processing yield maps from multiple sensors in large vineyards in California. In J. A. Taylor, D. Cammarano, A. Prashar, & A. Hamilton (Eds.), *Proceedings of the 11th European conference on precision agriculture*. *Advances in Animal Biosciences, 8,* 530–533.

Skøien, J. O., Baume, O., Pebesma, E. J., & Heuvelink, G. B. M. (2010). Identifying and removing heterogeneities between monitoring networks. *Environmetrics, 21,* 66–84.

Webster, R., & Oliver, M. A. (1992). Sample adequately to estimate variograms of soil properties. *Journal of Soil Science, 43,* 177–192.

Weller, U., Zipprich, M., Sommer, M., Castell, W. Z., & Wehrhan, M. (2007). Mapping clay content across boundaries at the landscape scale with electromagnetic induction. *Soil Science Society of America Journal, 71,* 1740–1747.