Investigating the harmonization of highly noisy heterogeneous datasets hand-collected over the same study domain

L. Pichon, C. Leroux, V. Geraudie, J. Taylor, Bruno Tisseyre

Investigating the harmonization of highly noisy heterogeneous datasets hand-collected over the same study domain.

Pichon, L.[1], Leroux, C.[1,2], Geraudie, V. [3], Taylor, J. [1], Tisseyre, B. [1]
[1]*ITAP, Univ Montpellier, Montpellier SupAgro, Irstea, , Montpellier, France*
[2]*SMAG, Montpellier, France*
[3]*Pellenc, Pertuis, France*
leo.pichon@supagro.fr

Abstract:
The objective of this paper is to propose an approach to harmonise noisy spatial data acquired by different operators using (low-cost) hand-held sensors over the same spatial domain. In such cases, datasets need to be harmonised before to be comparable before decision making. This work proposes a methodology to address this issue in the case of nested and noisy spatial data. First, it proposes the implementation of a non-parametric test of Kolmogorov-Smirnov to determine if harmonisation is needed. Then, it proposes an aspatial harmonization method based on a standardization. The method was applied on grape sugar content datasets collected by 2 hand-held spectrometers. Results showed that harmonizing a less confident dataset with respect to a more trustworthy one is interesting solely if the size of the trusted dataset is too small.

Keywords: Sugar content, spectrometer, Precision Viticulture, Crowdsourcing

Introduction

The use of hand-held measurement systems, particularly smartphones, is flourishing in agriculture. This is particularly true for perennial crops, such as vines, where many operations are still manual and data collection by operators with hand-held sensors is feasible (Aquino et al., 2018, Fuentes et al., 2012, Aquino et al., 2015, Geraudie et al. 2010). Spatial observations collected with such systems may be abundant and are following an increasing trend as the acquisition is simple and cost and time effective. In a production context, this permits the acquisition of multiple datasets by several operators simultaneously over the same spatial domain *D*, ranging from large (e.g. watershed) to small (e.g. sub-field/plot) domains, but sometimes under varying acquisition conditions, e.g., different operators, sensor model or calibration. These datasets have unique features:

i) Spatial observations are usually very noisy because measurements are performed on a small spatial support and the short range variability is often high on perennial crops (from one leaf/fruit to another or from one plant to another),
ii) The source of variability may be caused by the operator, who may be more or less aware of how the acquisition conditions affect the quality of the measurements. The resulting dataset quality may differ from one operator to another depending on the level of training, skill or attention paid to the calibration procedure, if such a procedure is necessary,
iii) The data collected by several operators over the same domain are generally intimately intertwined (nested). Operators effectively do not necessarily know the location of measurements performed by other operators.
iv) No reference dataset is generally available. In a production context, it is therefore not possible to assess the quality of data collected with hand-held measurement systems by comparing them to a reference.

The resulting datasets are not comparable and need to be harmonized before used in decision-making. While some research has addressed the general issue of data harmonization (Brenning et al., 2008, Sams et al., 2017), very few studies have directly considered the case of noisy spatial datasets collected over the same spatial domain *D*, such as agricultural crowdsourcing data (Minet et al. 2017). Some studies have proposed advanced data fusion approaches (Gé et al. 2014) but generally considering that reference data are available. This study proposed a methodology to address this specific issue. The proposed methodology was applied on real grape sugar content data collected by two different operators with a hand-held spectrometer each.

Materials and Methods

General approach

Consider several spatialized datasets (Y_1, Y_2,... Y_i) relating to the same agronomic phenomenon and collected by different operators using hand-held measurement systems on the same study domain *D* and with similar spatial distribution. These datasets are considered as estimations of the same regionalized variable (agronomic phenomenon) whose the first and second order moments is expected to be identical over the domain *D*. The set of conditions c_1, c_2… c_i under which these datasets were collected correspond to the characteristics of the operator (level of experience, level of attention, interest, etc.) and the sensor (calibration, model, etc.). The proposed approach assumes that there is at least one set of conditions corresponding to a properly calibrated sensor and to the best possible conditions of acquisition in a production context. The dataset acquired under these conditions is considered as a reference (or the best possible reference) even though its quality cannot be assessed properly.. In this paper, this set of conditions is noted c_1 and the resulting reliable dataset is noted Y_1. In order to avoid any confusion, Y1 will be referred to as best possible reference (BPR) in the rest of the document In a concern for simplicity, the rest of the approach is described for two datasets Y_1 and Y_2. Note however that, the approach remains transferable to *i* datasets (Y_1, Y_2,... Y_i).

The objective of this paper is to propose a simple harmonisation approach for highly noisy heterogeneous datasets hand-collected over the same study domain *D*. This paper aims at testing this simple approach on a real case study and to evaluate its relevance for a conjoint use of Y_1 and Y_2 in a decision support context. This relevance was evaluated through different scenarios.

Harmonization approach

First, the approach proposes to evaluate whether the conjoint use of Y_1 and Y_2 requires a pre-processing harmonization step. As a first approach a simple method based on attribute distribution was proposed. More sophisticated approach based on the covariance between Y1 and Y2 was not considered in this first approach mainly because of the noise and the number of available measurements which make it difficult to perform a proper spatial covariance analysis. Considering the common characteristics of Y_1 and Y_2 (same agronomic phenomenon observed, same domain *D* studied, similar spatial distribution), it is expected that their attribute distributions have to be similar. If not, it is considered necessary to harmonize Y_2 with Y_1. The Kolmogorov Smirnov statistical test (K–S test) was proposed to compare Y_2 and Y_1 distributions.. Harmonization was considered necessary when the null hypothesis is rejected ($p < 0.05$).

When harmonisation is necessary, it is considered that the acquisition conditions c_2 have generated a transformation of the attribute values of Y_2. This transformation function can be approximated by a function *f* such that:

$$Y_2 = f(Y_1) \qquad (1)$$

Depending on the conditions that generated this transformation, function f can be simulated in several different ways. As a first approach, it was considered that function *f* was a linear function (Eq. 2).

$$Y_2 = a.Y_1 + b \qquad (2)$$

In Eq. 2, the linear function makes it possible to account for the main sources of transformation of Y_2. The parameter a can be seen as a change in the sensitivity of the measuring system. It allows modelling changes in the sensor calibration or in the level of expertise and in operator's care when performing the measurement. Parameter b can be seen as a bias of the measuring system. It allows accounting for potential systematic bias due to the operator or sensor calibration.

The objective of the proposed harmonization approach was to estimate parameters a and b. As observations of Y_1 and Y_2 were not located on the same sites, a linear regression between these two datasets was not possible. In this respect, the following simple standardization method was proposed.

First, a was estimated as the value that could minimize the difference in variance between datasets Y_1 and Y_2 using a minimization function (Nelder and Mead, 1965). This parameter a was used to transform the dataset Y_2 into a new dataset Y_2' (Eq. 3).

$$Y_2' = a.Y_1 \qquad (3)$$

In a second step, parameter b was estimated as the value that could minimize the difference in mean value between Y_2' and Y_1 (centering) using the same minimization function. Parameter b was finally used to calculate $Y_{2_{harmonized}}$ (Eq. 4)

$$Y_{2_{harmonized}} = Y_2' + b \qquad (4)$$

Description of the case study

The use case corresponds to two datasets Y_1 and Y_2 of grape sugar content measurements collected using 2 portable spectrometers (Geraudie et al. 2010). Observations were collected in 2015 on a 0.5 ha vineyard block. The block was located about 20 km north of Aix en Provence in the Provence region of south of France. Observations were collected by 2 different operators. Operators have gone through the same 4 vine rows. Each operator randomly chose the bunches on which measurements were performed. Therefore, observations collected by the two different operators were not collocated in space. Grape sugar contents measured by the first operator, referred to as Y_1, were considered as the best possible reference. This choice may seem subjective, however in a production context, indicators (often well known) relating to the knowledge of operators, their level of training, their rigour in field measurements, etc. but also on the sensors they used (sensor model, renewal, calibration procedure, etc.) may be used to decide. . Y_1 and Y_2 both contain 200 observations.

Studied scenarios

In the study case, Y_1 and Y_2 contain the same number of observations. Nevertheless, in many real cases, Y_1 's observations are more complex and expensive to acquire than Y_2's. It is therefore common for Y_1 to contain fewer observations than Y_2. This is the case, for example, when Y_1 is collected by a trained operator using a reliable but expensive sensor and Y_2 is collected by many operators using a low-cost but less reliable sensor.

To tackle this issue, the sensitivity of the proposed approach to the diminution of the amount of data in Y_1 was tested. In this respect, the harmonization approach was tested with Y_{1_n} as the reference dataset. Y_{1_n} corresponds to Y_1 but with only a certain percentage of data from the original dataset, the rest being removed. "*n*" refers to this percentage, with "*n"* taking the values 10, 30, 50, 70 and 90. Five iterations were performed for each value of "*n*".

Three scenarios were considered. In scenario 1, the considered dataset was only Y_{1_n}. In scenario 2, it was considered Y_{1_n} to which was added the whole dataset Y_2 (without harmonization). In scenario 3, it was considered Y_{1_n} to which was added the whole dataset $Y_{2_{harmonized}}$.

Evaluation of the scenarios

Datasets of the different scenarios were used to produce sugar content maps following a block-kriging procedure on a regular 10 m square grid. Scenarios were evaluated by comparing these maps to the reference map. The RMSE was chosen as an indicator to summarise the observed pixel-to-pixel differences between both maps. Spatial observations were block-kriged in order to i) compare observations that were not collected at identical sites, ii) smooth the information of this relatively noisy signal and iii) work over a spatial support that is consistent with what is usually done in an operational context. This paper aims at comparing the quality of the estimation of the within-field sugar content for the different scenarios as a function of the quantity of data contained in Y_1. It is not intended to define an absolute prediction error for these scenarios. The objective was therefore to define a reference map that was relatively easy to build and that would allow scenarios to be compared to each other. The whole dataset Y_1 was chosen as a reference and the reference map was then built following the same block kriging procedure described above.

Results

Figure 1a shows that the spatial distribution of the two datasets (Y_1 and Y_2) are nested in space. The attribute distribution of Y_2 presents a very similar shape but biased when compared to the one of Y_1 (Fig.1b). Regarding their spatial structure, both Y_1 and Y_2 present a high erratic variance (nugget effect around 0.7) compared to their total variance (sill around 0.9).

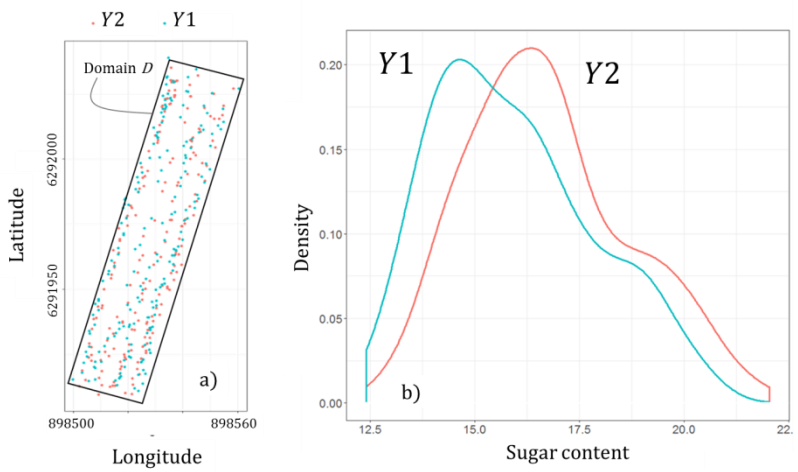

Figure 1: Spatial (a) and attribute (b) distribution of the two studied datasets

The application of the non-parametric Kolmogorov Smirnov test to Y_1 and Y_2 led to a p-value of 2.2×10^{-3}. Distributions are considered significantly different and the data harmonisation is required. The Kolmogorov-Smirnov test was considered to be the most relevant in the context of this study. However, particular attention should be paid in the case of small datasets (n < 30 ?) for which the p-value may be high even if the two datasets are different.

Figure 2 represents the RMSE of the sugar content prediction as a function of the size of the subset of Y_1 (from 10 to 90 % of data points). The red, blue and green boxplots represent respectively the RMSE of the five iterations of the predictions obtained from scenarios 1 (only Y_{1_n}), scenarios 2 (Y_{1_n}+ Y_2) and scenarios 3 (Y_{1_n}+ $Y_{2_{harmonized}}$).

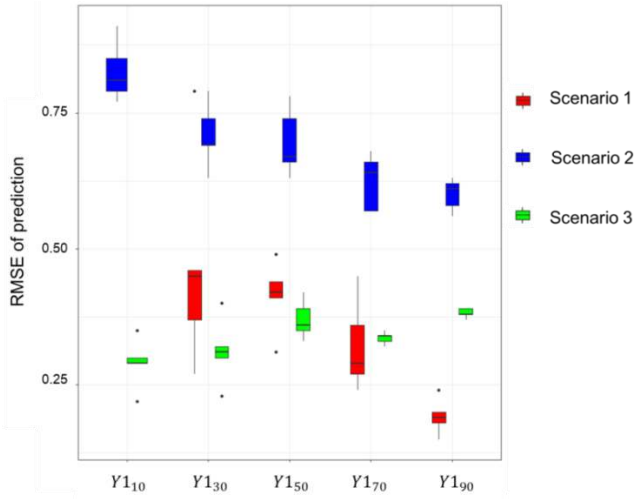

Figure 2: RMSE of prediction for the 3 scenarios depending on the quantity of data contained in Y_{1_n}

Note that no estimation was made for *n*=10 in scenario 1. The number of observations available (20 in this case) was too small for the variogram calculation to produce a map by block-kriging. For *n*=30, the number of observations available (60) was considered large enough. Regarding scenario 1, with an increasing *n*, the RMSE of the prediction decreased (Fig. 2). This result is not surprising and simply reflects the fact that as more

relevant data are considered, the estimation errors decrease. The RMSE of the predictions obtained in scenario 2 follows the same trend but with significantly higher RMSE values whatever the value of *n* (Fig. 2). The addition of the uncorrected (not harmonized) dataset Y_2 significantly increases the estimation errors in any case. This proves that there is no value, and in fact a cost, in merging both datasets with different confidence levels if those datasets are left as such, i.e. raw data. The RMSE obtained in scenario 3 remains relatively constant when *n* varies. The observed RMSE slightly oscillates between 0.3 and 0.4 (Fig. 2). This result can be explained by the fact that the $Y_{2_{harmonized}}$ dataset ensures a lower estimation error when the number of relevant data in Y_1 is small. In contrast, when the subset of Y_1 increases, the addition of the dataset $Y_{2_{harmonized}}$ may limit the quality of the prediction. The low RMSE value (around 0.2) observed with only $Y1_{n=90}$, is never reached (Fig. 2).

More generally, for high values of *n* (> 70%), the RMSE observed with Y_{1_n} (scenario 1) is always lower than that observed with Y_{1_n}+ Y_2 (scenario 3) (Fig. 2). This trend changes when *n* decreases and shows that there is a quantity of data in Y_1 below which the addition of the data set $Y_{2_{harmonized}}$ improves the quality of the resulting prediction. In this particular case, this amount of data corresponds to a value of 50% of Y_1. Therefore, when *n* is lower than this value (*n* = 50%), results show that it makes sense to use $Y_{2_{harmonized}}$ to produce a map of the grape sugar content.

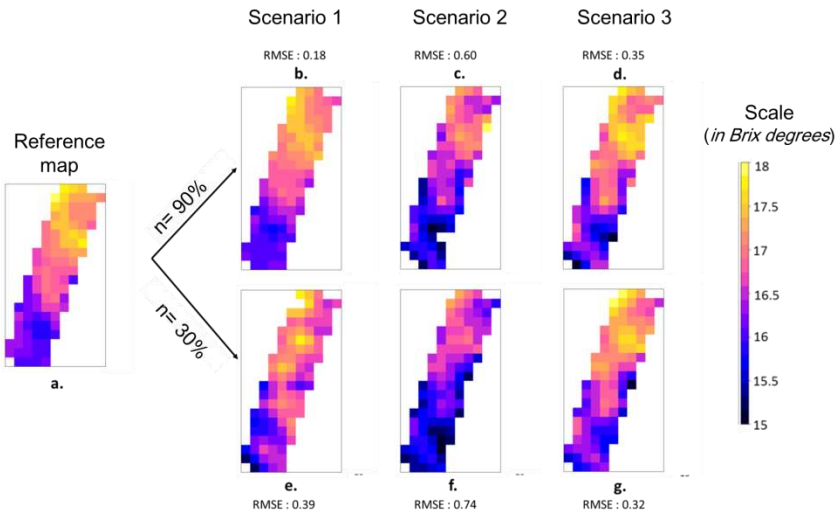

Figure 3: Reference map (a) and estimations maps produced for n=90% (b, c, d) and n=30% (e, f, g) according to the 3 scenarios

Figure 3 represents some of the block-kriged maps that were computed for different sizes of subsets of Y_1. For *n* equal to 90%, the spatial patterns obtained in scenario 1 (Fig. 3b) are the closest to those of the reference map (Fig. 3a). The southwestern/northeastern gradient of the sugar content is similar. There are also the three main zones, respectively low, medium and high sugar contents along this gradient. On the other hand, for *n* equal to 30%, the spatial patterns of scenario 1 are different from those of the reference. In this case, scenario 3 (Fig 3g) presents the closest spatial patterns to the reference. It also has the lowest RMSE (0.32 compared to 0.39 for scenario 1). In both cases (*n*=30% and *n*=90%), the spatial patterns of scenario 2 are the most different of the reference map.. The maps (Fig. 3) of the two example values of *n* (30% and 90%) confirm the results

presented in Figure 2: The use of the $Y_{2_{harmonized}}$ dataset improves the quality of sugar content maps only if the subset of Y_1 is small. In this study, and as a first approach, block-kriging was carried out with the same weight given to Y1 and Y2. In future work, it seems relevant, for scenarios 2 and 3, to be able to take into account the error variance of Y2 in the weight assigned to these data as described by Chiles and Delfiner (1999).
In this paper, the reference map corresponds to the entire dataset of Y_1. This is acknowledged to be a limitation. Ideally, the reference map should have been constructed from independent data, such as from laboratory measurements of berry sugar contents. The choice of this reference map is a first approach. It allows tackling this issue by taking into account the available data in an operational context. The resulting constraints do not allow the use of more traditional methods such as independent reference maps to be used. The choice of methods like cross validation or data splitting could have offered an alternative. However, they were not considered here mainly because the amount of data available would have been too small. Results must be therefore tempered, particularly for high *n* values. Indeed, when *n* increases, on the one hand Y_{1_n} tends towards Y_1. The RMSE of scenario 1 (only Y_{1_n}) then decreases towards 0. On the other hand, regarding scenario 3, $Y_{1_n} + Y_{2_{harmonized}}$ tends towards $Y_1 + Y_{2_{harmonized}}$. $Y_1 + Y_{2_{harmonized}}$ is different from the reference (Y_1) so the RMSE is higher than 0, whatever the quality of the data contained in $Y_{2_{harmonized}}$. This difference is due to the choice of the reference map. The differences in RMSE that are observed between scenarios 1 and 3 for *n* equals to 90% are therefore partly due to this choice and not only to the quality of the prediction. The choice of a better reference map would likely reduce the observed differences in RMSE. However, the trend would certainly be maintained.
This discussion raises the more global question of the choice of reference data. Indeed, this choice is often a major issue when evaluating the quality of datasets collected manually by different operators. When reference data are available, there are many methods available to assess the quality of the data collected (Flanagin and Metzger, 2008; Senaratne et al. 2016). On the other hand, in the literature, only a few papers have addressed the case where reference data were not available. Muller et al. (2015) provided an interesting approach by using the spatial and temporal coherence of observations (temperatures) to assess data quality. This approach seems particularly relevant to crowd-sourced datasets in the context of precision agriculture, where the studied phenomena are often spatially structured. Future work will also focus on the comparison of spatial and non-spatial harmonisation methods along with their advantages and disadvantages. In particular, attention should be paid to the sensitivity of these methods to the noise of the initial datasets.

Conclusion:

The development of hand-held measurement systems, and in particular smartphones, can lead to the collection of multiple datasets over the same spatial domain, but sometimes unfortunately under different acquisition conditions (operator, sensor, timing, etc...). It is therefore likely to generate nested and noisy datasets that require data harmonization before interpretation. This paper studied a simple harmonisation approach: standardisation. In this study, two datasets were compared and harmonized, one of which was considered much more trustworthy than the other. The proposed aspatial approach was tested on real grape sugar content datasets by varying the size of the confident dataset.

Results showed that harmonizing a less confident dataset with respect to a more trustworthy one is interesting solely if the size of the reliable dataset is too small.

Acknowledgements

We thank the Pellenc Company for making datasets available and for their expertise in their analysis.

References

Aquino, A., Barrio, I., Diago, M. P., Millan, B., & Tardaguila, J. (2018). vitisBerry: An Android-smartphone application to early evaluate the number of grapevine berries by means of image analysis. Computers and Electronics in Agriculture, 148, 19-28.

Aquino, A., Millan, B., Gaston, D., Diago, M. P., & Tardaguila, J. (2015). vitisFlower®: development and testing of a novel Android-smartphone application for assessing the number of grapevine flowers per inflorescence using artificial vision techniques. Sensors, 15(9), 21204-21218.

Brenning, A., Koszinski, S. & Sommer, M. (2008). Geostatistical homogenization of soil conductivity across field boundaries. Geoderma, 143(3–4), pp.254–260.

Chiles, J. and Delfiner, P. (1999) Geostatistics: Modelling Spatial Uncertainty. John Wiley, New York

Flanagin, A.J. & Metzger, M.J., 2008. The credibility of volunteered geographic information. GeoJournal, 72(3–4), 137–148.

Fuentes, S., De Bei, R., Pozo, C., & Tyerman, S. (2012). Development of a smartphone application to characterise temporal and spatial canopy architecture and leaf area index for grapevines. Wine & Viticulture Journal. J, 6, 56-60.

Ge, Y., Avitabile, V., Heuvelink, G.B.M., Wang, J & Herold, M. (2014), Fusion of pan-tropical biomass maps using weighted averaging and regional calibration data. International Journal of Applied Earth Observation and Geoinformation 31, 13 24

Geraudie, V., Roger, J.M. & Ojeda, H. (2010). Développement d'un appareil permettant de prédire la maturité du raisin par spectroscopie proche infra-rouge(PIR). (Development of a sensor to predict grape maturity by near infrared spectroscopy (NIR)). Revue Française d'Oenologie,. pp 2 - 8, <hal-00647105>

Minet, J., Curnel, Y., Gobin, A., Goffart, J.P., Mélard, F. & Tychon, B. (2017). Crowdsourcing for agricultural applications: A review of uses and opportunities for a farmsourcing approach. Computers and Electronics in Agriculture, 142, 126–138.

Muller, C.L., Chapman, L., Johnston, S., Kidd, C., Illingworth, S. & Foody, G. (2015). Crowdsourcing for climate and atmospheric sciences: Current status and future potential. International Journal of Climatology, 35(11), 3185–3203.

Nelder, J. A. & Mead, R. (1965). A simplex algorithm for function minimization. Computer Journal, 7, 308–313.

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

Senaratne, H., Mobasheri, A., Loai Ali, A., Capineri, C & Haklay, M. (2016). A review of volunteered geographic information quality assessment methods. International

Journal of Geographical Information Science, 31(1), 139–167.