DATA MINING APPROACHES FOR TIME SERIES DATA ANALYSIS IN VITICULTURE. POTENTIAL OF THE BLiSS (BAYESIAN FUNCTIONAL LINEAR REGRESSION WITH SPARSE STEP FUNCTIONS) METHOD TO IDENTIFY TEMPERATURE EFFECTS ON YIELD POTENTIAL

Authors:Cécile LAURENT[1,2,3,*] Meïli BARAGATTI[4], James TAYLOR[1], Bruno TISSEYRE[1], Aurélie METAY[2], Thibaut SCHOLASCH[3]

[1]*ITAP, Univ. Montpellier, Montpellier SupAgro, Irstea, France*
[2]*SYSTEM, Univ Montpellier, CIHEAM-IAMM, CIRAD, INRA, Montpellier SupAgro, France*
[3]*Fruition Sciences, Montpellier, France*
[4]*MISTEA, Univ Montpellier Montpellier SupAgro, INRA, France*

*Corresponding author : cecile@fruitionsciences.com

Abstract:

Context and purpose of the study – Vine development, and hence management, depends on dynamic factors (climate, soil moisture, cultural practices etc.) whose impact can vary depending upon their temporal modalities (timing, duration, threshold, eventually trajectory and memory effects). Therefore, understanding the effect of the temporal variation of these factors on grapevine physiology would be of strategic benefit in viticulture, for example in establishing yield potential. Today many estates own data that can support temporal analyses, while the emergence of precision viticulture allows management at higher spatial and temporal resolutions. These data are a great opportunity to advance knowledge about the dynamics of grapevine physiology and production, and promote an improved precision of vineyard practices. The exploitation of these data needs analytical methods that fully explore time series data. However, current methods tend to only focus on a few key phenological stages or time steps. Such approaches do not fully address the potential information captured by continuous temporal measurements because they introduce limitations : i) they rely on choices of variables and timing, ii) they often require suppressing data or analysing only parts of a time series and iii) data correlation over time is not taken into account. A new approach is explored in this paper, using a Bayesian functional Linear regression with Sparse Steps functions (BLiSS method). The BLiSS method overcomes the mentioned limitations and leads to a more complete and objective analysis of time series data. Based on the identification of climatic periods affecting yield, the objective of the study is to evaluate the potential of the BLiSS method.

Materials and method - Minimum and maximum daily temperatures during the year preceding the harvest year were regressed against the number of clusters per vine using the BLiSS method on one block of a commercial vineyard in the Bordeaux region over 11 years. The reliability and pertinence of the BLiSS method to reveal already reported, ignored or underestimated temperature effects on the number of clusters per vine are tested by comparison with literature results.

Results - The BLiSS method allowed the detection of periods when temperature influenced the number of clusters per vine during the year preceding the harvest year. Some of the detected periods of influence had already been reported in literature. However, the BLiSS outcomes suggested that some of those known periods may have a different duration or several effects, thus challenging actual knowledge. Finally, some new periods of influence were identified by the BLiSS method. These results confirmed the potential of the BLiSS method to undertake a fuller exploration of time series data in the case of climate influence on grape yield.

Keywords: climate, functional analysis, temporal variability, cluster number

1. Introduction

Understanding precisely the temporal progression of physiological processes is an important step to improve any crop management. It helps to refine decision-making and the scheduling of pertinent field observations that are adapted to seasonal conditions. Due to its perennial nature, vine physiological processes are interdependent at intra- and interannual levels. This interdependency sometimes results in complex autoregulation, memory or trajectory effects (Vaillant-Gaveau et al., 2014). Understanding the temporal progression of a vines physiological process is therefore complicated by the temporal interdependencies that need to be taken into account.

The temporal analysis of a physiological process requires the identification of both periods and variables that correspond to an increased sensitivity in the response. Once potentially influential periods for a given variable have been identified, a second analysis can be performed in order to investigate the explanatory mechanisms that lay behind these detected influences. To the best of the authors' knowledge, aside from field expertise, there are two main approaches to determine the combinations of variables and periods that are influential on a physiological process. The first approach consists of field or laboratory experiments, which focuses on a few hypothesized periods and influencing variables (Buttrose, 1974 ; Pagay & Collins, 2017). However, this approach only allows a few combinations of critical stages and influencing variables to be tested, which is its main limitation. The second approach, based on data mining, seeks to address such limitations. However, to perform these analysis easily with existing statistical methods, such as linear regression, most studies have segmented an influencing variable time series (*i.e.* history) into synthetic sub-variables, for example the sum of rain over 10 days (Guilpart et al., 2014 ; Molitor and Keller, 2017). This simplification reduces the detection of critical periods to the chosen time step and often focuses the analysis over a fraction of the data time series. Therefore, information that could improve the understanding of crop physiological processes is potentially missed. Moreover, the sub-variables are independently analysed using these classical methods. As such, the potential influence of data correlation over time is not taken into account. For example, the impact of summer rain should be assessed relative to winter and spring rain levels. Therefore, this second approach can be misleading as it may highlight influencing periods without considering the historical context.

With changing data structures, new methods are needed to fully explore data time series and improve understanding of physiological processes, particularly in perennial crops. As a case study, this paper evaluates the potentialities of a Bayesian functional Linear regression with Sparse Steps functions (Grollemund et al., 2019), the BLiSS method, in detecting influencing periods and variables on a given quantitative physiological indicator. The BLiSS method allows a time series to be fully analysed while considering data correlation over time. The continuous effect of temperature in year *n-1* on cluster number per vine (CN) in year *n* is investigated. Grapevine yield is determined over two years, the year of harvest (year *n*)and the previous year (year *n-1*). CN is considered to be the main yield-determining component, affecting 60% of final yield variability, and is mainly determined during year *n-1* (Guilpart et al., 2014, Li-Mallet et al., 2016). In this way, understanding which variables and periods affect the CN elaboration process could help early yield forecasting. Temperature during year *n-1* is known to be highly influential on bud fruitfulness, which directly impacts CN (Li-Mallet et al., 2016). Beyond yield parameters, such as CN, numerous papers have emphasised the need to analyse time series to highlight the effects of environmental variables on other vine parameters, such as phenological stages (Zapata et al, 2017) or berry parameters (Rienth et al., 2014 ; Pastore et al., 2017). Other recent studies have also challenged the actual knowledge about the influence duration of some environmental variables (Gaiotti et al., 2018). This paper aims to assess the potentialities of the BLiSS method to highlight periods of significant temperature influence on CN and validate its relevance to explore times series in any similar cases.

2. Material and Methods

Data from a commercial Cabernet-Sauvignon (*Vitis vinifera* cv. Cabernet Sauvignon) vineyard field (~1 ha) in the Bordeaux region (France) were used for the study. CN was recorded after fruitset on the same 100 vines from 2007 to 2018. For the BLiSS analysis, each individual corresponds to a CN observation in each year. Daily minimum (*Tmin*) and maximum (*Tmax*) temperature (°C) were collected by a local weather station. To reduce processing time, temperature data were aggregated into 10-day periods. For the BLiSS analysis, each individual was regressed against the *Tmin* and *Tmax* time series from March 1st to October 25th of year *n-1*.

The analysis was performed using the package bliss version 1.0.0 (Grollemund, 2019) in R 3.5.1 (R Core Team, 2018). The *Tmin* and *Tmax* times series were analysed independently.

3. Results and discussion

The four detected periods of *Tmin* influence during year *n-1* are presented on Fig.1(A). Periods (i), (ii) and (iv) are positively correlated to CN *i.e.* high *Tmin* occurring during these periods increased CN, whereas period (iii) is negatively correlated to CN *i.e.* high *Tmin* during this period decreased CN. Periods (i) and (iii) are characterized by a lower intensity but higher degree of confidence than periods (ii) and (iv) *i.e.* the influence of periods (i) and (iii) on CN might be of lower intensity but occurs more frequently, whereas the influence of periods (ii) and (iv) might be of higher intensity but happens over fewer seasons. The duration of periods (i) and (iii) exceeds one month, whereas periods (ii) and (iv) last approximately three weeks. Similarly, the three periods of *Tmax* influence on CN are presented in Fig.1(B). Periods (v) and (vii) are positively correlated to CN whereas period (vi) is negatively correlated to CN. Period (v) is characterized by two modalities : one related to a two-month period of low impact with a high degree of confidence and one related to a one-month period (around May) of higher impact but a medium degree of confidence. It means that in this two-month period, *Tmax* moderately increases CN over the duration of the period more often than it does with higher intensity for a short time in the middle of the period. Figures 1(A) and 1(B) periods are mainly characterized by similar degrees of confidence but mostly differ in duration and intensity.

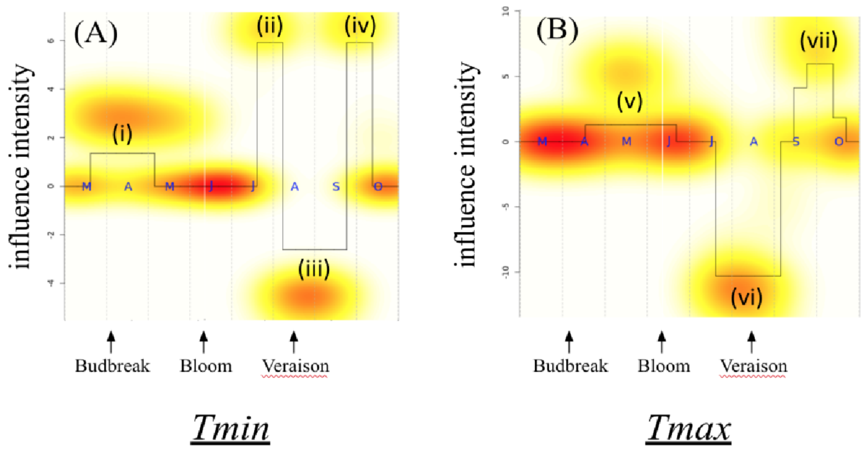

Figure 1. Graphical representation of the BLiSS output showing possible periods of influence respectively of (A) Tmin and (B) Tmax on CN from March 1st to October 25th of year n-1. The black line indicates the duration and the intensity of the influence over time. The colour gradient from yellow to red illustrates the degree of confidence for the detected period, respectively from low to high. The x axis labels M,A,M,J,J,A,S,O represent March, April, May, June, July, August, September, October respectively. Average dates of budbreak, bloom and veraison recorded from 2006 to 2018 are indicated.

3.1 Agreement between the BLiSS results and reported literature

Periods (i), (v), (ii), (vi) and (iii) respectively coincided with budbreak, bloom, veraison and maturation periods. The detected periods for temperature influence are in agreement with literature results (Guilpart et al., 2014 ; Li-Mallet et al., 2016 ; Molitor & Keller, 2017) but sometimes suggest a different duration or the existence of several effects. These results confirm that the BLiSS method can effectively detect periods of temperature influence on CN that are already known and previously reported in literature. Periods (iv) and (vii) were not reported in literature but detected with a medium degree of confidence by the BLiSS method. They refer to a post-harvest period during year *n-1*. The method sheds new light on the duration of known periods of influence and suggests the existence of previously unreported periods influencing CN. These results do need to be validated with viticulture experiments.

3.2 The BLiSS method allows an exploratory analysis of time series

The BLiSS method allowed an objective detection of periods of temperature influence on CN during the year *n-1.* Each detected period was characterized by a duration, an intensity and a degree of confidence. However, further understanding of temperature effects on CN determination in year *n-1* will require classifying the detected periods by order of decreasing influential importance. Such a classification requires balancing the criteria of duration, intensity and degree of confidence according to one's objectives. This can be done in a second step, either by introducing expert knowledge or by using explanatory methods that would involve the periods that have already been detected.

4. Conclusion

The ability of the BLiSS method, as an exploratory approach, to reveal new potential periods of influence on vine physiological processes and to challenge actual knowledge about the duration of these influential periods shows great promise. In addition, detecting in parallel analysis the influential periods for the same variable on different vine physiological parameters could also help in identifying some synergies and possible explanatory mechanisms of vine physiology. Data mining approaches, such as the BLiSS method, could allow a more pertinent time series exploration prior to any work using ecophysiological, genetical, chemical or other indicators.

5. Acknowledgements

The authors fully acknowledge the commercial estate that provided the data as well as the authors of the BLiSS method who spent time explaining its application.

6. Literature cited

BUTTROSE M., 1974. Climatic factors and fruitfulness in grapevines. Horticultural Abstracts 44, 319–326.

GAIOTTI, F., PASTORE, C., FILIPPETTI I., LOVAT L., BELFIORE N., TOMASI D., 2018. Low night temperature at veraison enhances the accumulation of anthocyanins in Corvina grapes (Vitis Vinifera L.). Scientific Reports 8, 8719.

GROLLEMUND P.-M., ABRAHAM C., BARAGATTI M., PUDLO P., 2019. Bayesian Functional Linear Regression with Sparse Step Functions. Bayesian Anal. 14, 111–135.

GUILPART N., METAY A., GARY C., 2014. Grapevine bud fertility and number of berries per bunch are determined by water and nitrogen stress around flowering in the previous year. European Journal of Agronomy 54, 9–20.

LI-MALLET A., RABOT A., GENY L., 2016. Factors controlling inflorescence primordia formation of grapevine: their role in latent bud fruitfulness? A review. Botany 94, 147–163.

MOLITOR D., KELLER, M., 2017. Yield of Müller-Thurgau and Riesling grapevines is altered by meteorological conditions in the current and previous growing seasons. OENO One 50, 245–258.

PAGAY V., COLLINS C., 2017. Effects of timing and intensity of elevated temperatures on reproductive development of field-grown Shiraz grapevines. OENO One 51, 4, 409-421.

PASTORE C., DAL SANTO S., ZENONI S., MOVAHED N., ALLEGRO G., VALENTINI G., FILIPPETTI I., TORNIELLI G.B., 2017. Whole Plant Temperature Manipulation Affects Flavonoid Metabolism and the Transcriptome of Grapevine Berries. Front Plant Sci 8, 929.

RIENTH, M., TORREGROSA L., LUCHAIRE N., CHATBANYONG R., LECOURIEUX D., KELLY M.T., ROMIEU C., 2014. Day and night heat stress trigger different transcriptomic responses in green and ripening grapevine (vitis vinifera) fruit. BMC Plant Biology 14, 108.

VAILLANT-GAVEAU N., WOJNAROWIEZ G., PETIT A.-N., JACQUENS L., PANIGAI L., CLEMENT C., FONTAINE F., 2014. Relationships between carbohydrates and reproductive development in chardonnay grapevine: impact of defoliation and fruit removal treatments during four successive growing seasons. OENO One 48, 219–229.

R CORE TEAM, 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

ZAPATA D., SALAZAR-GUTIERREZ M., CHAVES B., KELLER M., HOOGENBOOM G., 2017. Predicting Key Phenological Stages for 17 Grapevine Cultivars (Vitis vinifera L.). Am J Enol Vitic. 68, 60–72.