

# 2526-2-F1801Q161 Causal Networks

-

## Practical Exam Instructions

Alessio Zanga<sup>1,\*</sup>

Models and Algorithms for Data and Text Mining Laboratory (MADLab),  
Department of Informatics, Systems and Communication (DISCo),  
University of Milano-Bicocca, Milan, Italy

### 1 Tasks

In the following pages you can find the paper titled:

Magrini, Alessandro, Stefano Di Blasi, and Federico Stefanini.  
"A conditional linear Gaussian network to assess the  
impact of several agronomic settings on the quality of  
Tuscan Sangiovese grapes." Biometrical Letters 54.1 (2017).

This paper describes a Conditional Linear Gaussian Network for canopy management techniques of Tuscan Sangiovese grapes, where the **Treatment** node encodes a categorical distribution with multiple treatments, while other nodes are multivariate Gaussian distributions. Assume the network is causal.

The objective of this exam is to compute the following counterfactual queries, where **T** is the **Treatment**:

– Total effects:

$$Q_1: \mathbb{E} [\text{Anthoc}_{(T = T4a)} - \text{Anthoc}_{(T = T1a)}]$$

$$Q_2: \mathbb{E} [\text{Brix}_{(T = T4a)} - \text{Brix}_{(T = T1a)}]$$

– Natural direct effects:

$$Q_3: \mathbb{E} [\text{Polyph}_{(T = T4a, \text{BunchN}_{(T = T1a)})} - \text{Polyph}_{(T = T1a, \text{BunchN}_{(T = T1a)})}]$$

$$Q_4: \mathbb{E} [\text{Polyph}_{(T = T4a, \text{Anthoc}_{(T = T1a)})} - \text{Polyph}_{(T = T1a, \text{Anthoc}_{(T = T1a)})}]$$

– Natural indirect effects:

$$Q_5: \mathbb{E} [\text{pH}_{(T = T1a, \text{BunchN}_{(T = T4a)})} - \text{pH}_{(T = T1a, \text{BunchN}_{(T = T1a)})}]$$

$$Q_6: \mathbb{E} [\text{GrapeW}_{(T = T1a, \text{Anthoc}_{(T = T4a)})} - \text{GrapeW}_{(T = T1a, \text{Anthoc}_{(T = T1a)})}]$$

---

\* Corresponding author: [alessio.zanga@unimib.it](mailto:alessio.zanga@unimib.it)

## 2 Instructions

You are given two files:

- `sangiovese.gml` - The graph of the model depicted in Figure 1, and
- `sangiovese.csv` - A dataset generated from the model.

You are free to choose any estimation procedure of your choice, as long as you provide a short report containing, for each query, the following:

- A natural language translation of the queries,
- A description of the identification process,
- A description of the estimation process,
- A natural language explanation of the estimates.
- The source code of the entire pipeline.

For simplicity, you can submit a single Jupyter Notebook in `.ipynb` format with all the required information.

You can use the packages we presented during the practicals, e.g. `networkx`, `dowhy`, `pgmpy`, and other existing packages, or write your own code.

You can find deadlines and additional delivery instructions in the **Exam Instructions** file in the **Exam** section of the e-learning platform.

## **A conditional linear Gaussian network to assess the impact of several agronomic settings on the quality of Tuscan Sangiovese grapes**

**Alessandro Magrini<sup>1</sup>, Stefano Di Blasi<sup>2</sup>, Federico Mattia Stefanini<sup>1</sup>**

<sup>1</sup>Department of Statistics, Computer Science, Applications – University of Florence, Florence, Italy, corresponding author. e-mail: stefanini@disia.unifi.it

<sup>2</sup>R&D wine and sensory consultant – Marchesi Antinori, Florence, Italy

### **SUMMARY**

In this paper, a Conditional Linear Gaussian Network (CLGN) model is built for a two-year experiment on Tuscan Sangiovese grapes involving canopy management techniques (number of buds, defoliation and bunch thinning) and harvest time (technological and late harvest). We found that the impact of the considered treatments on the color of wine can be predicted still in the vegetative season of the grapevine; the best treatments to obtain wines with good structure are those with a low number of buds; the best treatments to obtain fresh wines suitable for young consumers are those with technological rather than late harvest, preferably with a high number of buds, and anyway with both defoliation and bunch thinning not performed.

**Key words:** Canopy management; Conditional independence; Directed acyclic graphs; Late grape harvest; Polyphenolic content; Potential alcohol.

### **1. Introduction**

The dependence of the quality of wine on the quality of grapes is by now a consolidated area of knowledge. Grapes with excellent color (high anthocyanin content) and good structure (high polyphenolic content) make it possible to obtain a wide variety of wines (Ribereau-Gayon et al., 1999), while grapes with moderate alcoholic gradation and good acidic balance may be new oenological objectives (Kontoudakis et al., 2011). Consorzio Tuscania performed an experiment on two Sangiovese vineyards in Tuscany, Italy, with the purpose of studying the role played by canopy management techniques (number of buds,

defoliation and bunch thinning) and harvest time (technological and late harvest) in determining the quality of grapes.

In this paper, a Conditional Linear Gaussian Network (CLGN) model is built to relate canopy management techniques and harvest time with multiple outcomes representing the quality of grapes. CLGNs were firstly proposed by Lauritzen and Wermuth (1989), but the core idea may be traced back to path analysis (Wright, 1934). In a CLGN, a multivariate Gaussian distribution on multiple outcomes is assumed conditionally to treatments, and it is factored into linear regression models. In this way, possible conditional independence relationships among the outcomes are made explicit. By discovering conditional independence relationships before estimating the impact of treatments on the multiple outcomes, the number of parameters representing the multivariate Gaussian distribution is reduced, thus entailing more efficient estimates. Also, conditional independence relationships define the relevant predictors for the impact of treatments on each single outcome: once the values of the relevant predictors are known, no other information is required to predict the impact of treatments on that outcome.

The paper is structured as follows. In section 2, materials and methods are detailed. Results are reported in section 3, while section 4 contains a discussion of the findings.

## **2. Materials and methods**

In this section, we describe the experimental setting (subsection 2.1), data (2.2), and the statistical model used to perform the analysis (2.3).

### **2.1. Experimental setting**

Experiments took place on two Sangiovese vineyards in Tuscany: Brolio (Chianti Classico) and Le Mortelle (Monteregio di Massa Marittima). The vineyards had homogenous age, planting density and growth habits. Four blocks of different vegetative vigour were chosen in each vineyard, and to each block, eight

treatments were applied, derived from a combination of different canopy management techniques:

- number of buds: one bud (low), or three buds (high);
- defoliation: not performed, or performed at 50%;
- bunch thinning: not performed, or performed at 50%.

The Edmund Mach Foundation (IASMA), Trento, Italy, gathered data in 2007, 2008 and 2009. After harvesting, two groups of five plants were randomly chosen from each block. Then, before the next harvest, the following measurements were performed on each group:

- weight of wood (WoodW);
- mean number of sprouts (SproutN);
- mean number of bunches (BunchN);
- mean weight of grapes (GrapeW);
- Soil-Plant Analysis Development in June (SPAD06) and in August (SPAD08);
- Normalized Difference Vegetation Index in June (NDVI06) and in August (NDVI08).

The result of each of these measures was averaged for each group of plants. In autumn, technological harvesting was applied to the first group to form a first must, and late harvesting was applied to the other group to form a second must. The following measurements were performed on each must:

- total acidity (Acid);
- potassium content (Potass);
- potential alcohol (Brix);
- pH (pH);
- total anthocyanin content (Anthoc);
- total polyphenolic content (Polyph).

The data consist of 48 statistical units (two groups of plants in four blocks in two vineyards for three years). For simplicity, we combined the three experimental factors with the two harvest times, thus obtaining a total of 16

treatments (Table 1). The reference treatment is the one characterized by low number of buds, no defoliation, no bunch thinning and technological harvest.

**Table 1.** Treatments. ‘T1a’ is the reference

| Treatment code | Number of buds | Defoliation      | Bunch thinning   | Harvest time  |
|----------------|----------------|------------------|------------------|---------------|
| T1a            | Low            | Not performed    | Not performed    | Technological |
| T1b            | Low            | Not performer    | Not performed    | Late          |
| T2a            | Low            | Not performed    | Performed at 50% | Technological |
| T2b            | Low            | Not performer    | Performed at 50% | Late          |
| T3a            | Low            | Performed at 50% | Not performed    | Technological |
| T3b            | Low            | Performed at 50% | Not performed    | Late          |
| T4a            | Low            | Performed at 50% | Performed at 50% | Technological |
| T4b            | Low            | Performed at 50% | Performed at 50% | Late          |
| T5a            | High           | Not performed    | Not performed    | Technological |
| T5b            | High           | Not performer    | Not performed    | Late          |
| T6a            | High           | Not performed    | Performed at 50% | Technological |
| T6b            | High           | Not performer    | Performed at 50% | Late          |
| T7a            | High           | Performed at 50% | Not performed    | Technological |
| T7b            | High           | Performed at 50% | Not performed    | Late          |
| T8a            | High           | Performed at 50% | Performed at 50% | Technological |
| T8b            | High           | Performed at 50% | Performed at 50% | Late          |

## 2.2. Statistical model

A Conditional Linear Gaussian Network (CLGN: Lauritzen and Wermuth, 1989) is defined on a set of variables  $\mathbf{X} = \{X_1, \dots, X_p\}$ , each being either continuous with domain on the real numbers or qualitative with a finite number of values, and is composed of:

1. a *qualitative part*, encoded by a directed acyclic graph (DAG: Lauritzen, 1999), showing the factorization of the joint probability distribution of variables in  $\mathbf{X}$ :

$$P(\mathbf{X}) = \sum_{j=1}^p P(X_j | \Pi_j)$$

such that the univariate probability distribution of variable  $X_j$  ( $j = 1, \dots, p$ ) is conditioned to a vector of variables  $\mathbf{\Pi}_j$ , which cannot include continuous variables if  $X_j$  is a qualitative variable. In the DAG, each variable is represented by a node, a node receives a directed edge from another node if the univariate probability distribution of the variable represented by the former is conditioned to the variable represented by the latter, and no directed cycles are present. The qualitative part implies a set of conditional independence statements among the variables, which can be read off the DAG using specific rules (Lauritzen et al., 1990);

2. a *quantitative part*, that is a statistical model for each variable  $X_j$  ( $j = 1, \dots, p$ ) in the joint probability distribution  $P(\mathbf{X})$ :

2.1. if  $X_j$  is a continuous variable,  $P(X_j | \mathbf{\Pi}_j)$  is a linear regression model:

$$X_j | \mathbf{\Pi}_j = \boldsymbol{\pi}_j \sim N \left( \begin{pmatrix} 1 \\ \boldsymbol{\pi}_j \end{pmatrix}' \boldsymbol{\beta}_j, \sigma_j^2 \right)$$

where  $\boldsymbol{\pi}_j$  is a vector belonging to the joint sample space of variables in  $\mathbf{\Pi}_j$ , and  $\boldsymbol{\beta}_j$  is the vector of regression coefficients;

- 2.2. otherwise,  $P(X_j | \mathbf{\Pi}_j)$  is a conditional probability table, that is a set of discrete probability distributions of  $X_j$ , one for each configuration of variables taken by vector  $\mathbf{\Pi}_j$ .

The measurements described in 2.1 were included in the CLGN as continuous variables, together with a qualitative variable representing the 16 treatments (Table 1). Since measurements are strictly positive and they took place in different years, the logarithmic transformation was applied to map their sample space to the  $p$ -dimensional reals  $\mathbb{R}^p$ , and the annual mean was subtracted from each datum in order to eliminate annual heterogeneity. Prior constraints on edges were applied on the basis of causal knowledge: edges not respecting the temporal order in Table 2 were forbidden, and the following edges were forced to be present: Acid  $\rightarrow$  pH, Potass  $\rightarrow$  pH, Brix  $\rightarrow$  GrapeW, Anthoc  $\rightarrow$  Polyph. Details on the relation between conditional independence and causality can be found in Pearl (2009). Given these constraints, we applied a greedy search procedure on the space of possible DAGs based on the Bayesian Information Criterion, and the

quantitative part of the resulting DAG was estimated by maximizing the likelihood function. All computations were performed in R 3.3.2 for Windows (R Core Team, 2016) using the package bnlearn (Scutari, 2010).

**Table 2.** Temporal order. Outcomes in layers with lower (higher) index are temporally precedent (subsequent), while outcomes at the same layer are contemporary

---

|    |                                                |
|----|------------------------------------------------|
| 1. | SproutN                                        |
| 2. | NDVI06, SPAD06                                 |
| 3. | BunchN                                         |
| 4. | SPAD08, NDVI08                                 |
| 5. | WoodW                                          |
| 6. | GrapeW, Acid, Potass, Brix, pH, Anthoc, Polyph |

---

### 3. Results

The resulting DAG is shown in Figure 1. This DAG has 50 fewer edges than the maximal one, thus 50 parameters are saved with respect to an unrestricted multivariate Gaussian model. A summary of parameter estimation is reported in the Appendix. Relevant predictors and estimates of the impact of treatments on each outcome are reported in 3.1 and 3.2 respectively.

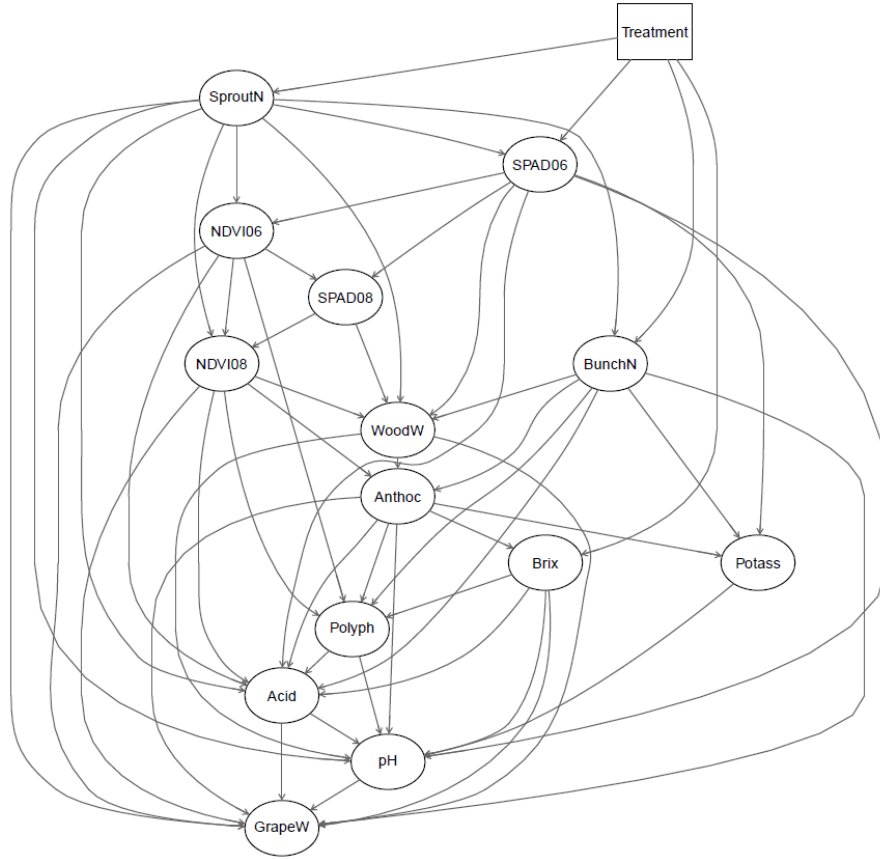
#### 3.1. Relevant predictors for the impact of treatments

Table 3 reports the minimal sets of outcomes making each outcome independent of treatments, found using the algorithm proposed by Tian et al. (1998). Each of these sets represents the relevant predictors for the impact of treatments on an outcome: once the values of outcomes in the set are known, no other information is required to predict the impact of treatments on that outcome.

The results show that the number of sprouts, the number of bunches, SPAD measured in June and potential alcohol are directly related to treatments in the DAG; that is, no outcome exists making them independent of treatments, and thus the impact of treatments on them cannot be predicted indirectly by the impact of treatments on other outcomes. However, the impact of treatments on the



anthocyanin content can be predicted by the impact of treatments on the number of bunches, NDVI measured in August and the weight of wood. This result is quite interesting, as it suggests that the impact of the considered treatments on the anthocyanin content can already be predicted in the vegetative season.



**Figure 1.** The DAG resulting from the greedy search procedure based on the Bayesian Information Criterion. The box-shaped node representing treatments denotes the absence of an observational distribution

**Table 3.** The minimal set of outcomes making each outcome independent of treatments

|         |                                                                 |
|---------|-----------------------------------------------------------------|
| SproutN | –                                                               |
| BunchN  | –                                                               |
| GrapeW  | Acid, Anthoc, Brix, BunchN, NDVI06, NDVI08, pH, SproutN, WoodW. |
| WoodW   | BunchN, SPAD06, SproutN.                                        |
| SPAD06  | –                                                               |
| NDVI06  | SPAD06, SproutN.                                                |
| SPAD08  | NDVI06, SPAD06.                                                 |
| NDVI08  | NDVI06, SPAD08, SproutN.                                        |
| Acid    | Anthoc, Brix, BunchN, SPAD06, SproutN.                          |
| Potass  | Anthoc, BunchN, SPAD06.                                         |
| Brix    | –                                                               |
| pH      | Acid, Anthoc, Brix, Polyph, Potass, SPAD06, SproutN, WoodW.     |
| Anthoc  | BunchN, NDVI08, WoodW.                                          |
| Polyph  | Anthoc, Brix, BunchN, NDVI06, NDVI08.                           |

### 3.2. Impact of treatments

For each treatment  $t_k$  ( $k = 0, 1, \dots$ ), our CLGN model implies a multivariate Gaussian distribution on the logarithm of multiple outcomes:

$$\log \mathbf{X} | t_k \sim MVN_p(\boldsymbol{\mu}_k, \Sigma) \quad k = 0, 1, \dots$$

We define the impact of a non-reference treatment on multiple outcomes as the ratio:

$$\boldsymbol{\rho}_k = \frac{E[\mathbf{X} | t_k]}{E[\mathbf{X} | t_0]} = \frac{\exp(\boldsymbol{\mu}_k + \text{diag}(\Sigma)/2)}{\exp(\boldsymbol{\mu}_0 + \text{diag}(\Sigma)/2)} = \exp(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)$$

where  $t_0$  is the reference treatment. Vectors  $\boldsymbol{\mu}_k$  ( $k = 0, 1, \dots$ ) can be obtained by recursively applying the formula:

$$E[\log X_j | t_k] = (1, E[\log \boldsymbol{\pi}_j | t_k])' \boldsymbol{\beta}_j \quad j = 1, \dots, p$$

The quantity  $(\boldsymbol{\rho}_k - 1) \cdot 100$  is the percentage variation in the expected value of multiple outcomes when  $t_k$  is applied rather than  $t_0$ . Table 4 shows the estimated

impacts of each non-reference treatment on the multiple outcomes predicted by the CLGN model, expressed as percentage variations.

**Table 4.** Estimated impacts of treatments on the multiple outcomes predicted by the CLGN model, expressed as percentage variations

| %       | T1b   | T2a    | T2b    | T3a    | T3b    | T4a    | T4b    |
|---------|-------|--------|--------|--------|--------|--------|--------|
| SproutN | +1.22 | +0.03  | +1.33  | +2.32  | +4.86  | +3.33  | +7.70  |
| BunchN  | +4.79 | -19.99 | -25.32 | +0.02  | +8.53  | -28.21 | -22.72 |
| GrapeW  | -8.65 | -20.85 | -31.30 | -11.57 | -13.77 | -33.23 | -35.44 |
| WoodW   | +1.40 | -3.68  | -0.46  | -5.75  | -0.95  | -10.81 | -5.57  |
| SPAD06  | +0.30 | -0.71  | +0.87  | -2.51  | -1.29  | -3.43  | -2.08  |
| NDVI06  | +0.23 | -0.25  | +0.47  | -0.89  | -0.28  | -1.34  | -0.50  |
| SPAD08  | +0.15 | -0.57  | +0.76  | -1.96  | -0.95  | -2.80  | -1.53  |
| NDVI08  | +0.23 | -0.35  | +0.37  | -0.85  | -0.01  | -1.11  | +0.05  |
| Acid    | -4.52 | -3.01  | -8.30  | -1.59  | -5.88  | -3.98  | -8.58  |
| Potass  | -0.26 | +1.61  | +2.77  | -0.95  | -1.08  | +1.64  | +1.59  |
| Brix    | +8.55 | +2.84  | +10.39 | +5.16  | +12.91 | +5.35  | +13.51 |
| pH      | +2.27 | +0.99  | +3.40  | +0.96  | +3.14  | +1.40  | +3.76  |
| Anthoc  | -1.48 | +4.50  | +3.84  | +2.24  | -0.67  | +8.93  | +5.36  |
| Polyph  | +1.67 | +2.07  | +3.14  | +2.82  | +3.56  | +4.73  | +5.16  |

| %       | T5a    | T5b    | T6a    | T6b    | T7a    | T7b    | T8a    | T8b    |
|---------|--------|--------|--------|--------|--------|--------|--------|--------|
| SproutN | +23.30 | +32.50 | +22.14 | +29.90 | +31.17 | +28.76 | +31.02 | +34.09 |
| BunchN  | +34.59 | +39.79 | -9.05  | -7.19  | +28.50 | +30.80 | -7.57  | -11.07 |
| GrapeW  | +24.27 | +15.69 | -11.78 | -17.20 | +10.12 | +3.28  | -15.74 | -25.84 |
| WoodW   | +3.50  | +5.98  | +2.61  | +6.52  | +1.24  | +4.58  | +5.58  | +5.37  |
| SPAD06  | -2.37  | -2.40  | -0.97  | -0.38  | -3.84  | -2.44  | -0.97  | -1.07  |
| NDVI06  | +0.12  | +0.48  | +0.65  | +1.27  | -0.18  | +0.33  | +1.12  | +1.15  |
| SPAD08  | -1.58  | -1.51  | -0.47  | +0.13  | -2.66  | -1.52  | -0.28  | -0.38  |
| NDVI08  | +1.44  | +2.29  | +1.78  | +2.89  | +1.51  | +1.97  | +2.91  | +3.02  |
| Acid    | +1.60  | -2.47  | -3.96  | -8.02  | -0.75  | -4.77  | -6.87  | -10.25 |
| Potass  | -3.39  | -3.74  | +0.22  | +0.13  | -3.57  | -3.15  | +0.18  | +0.25  |
| Brix    | -1.11  | +5.34  | +2.21  | +8.47  | +2.68  | +9.57  | +6.28  | +12.69 |
| pH      | -0.39  | +1.51  | +1.33  | +3.28  | +0.64  | +2.61  | +2.69  | +4.39  |
| Anthoc  | -5.74  | -7.23  | -0.47  | -2.39  | -4.53  | -6.10  | -2.01  | -1.32  |
| Polyph  | -1.68  | -0.56  | -0.14  | +0.59  | -0.15  | +1.08  | +0.34  | +2.42  |

The weight of the grapes is certainly decreased by treatments, unless the number of buds is high and bunch thinning is not performed (T5a, T5b, T7a, T7b). Reasonably, if the number of buds is low, the reduction is stronger in treatments including bunch thinning (T2a, T2b, T4a, T4b).

As largely expected, the number of bunches is certainly decreased by all of the treatments including bunch thinning (T2a, T2b, T4a, T4b, T6a, T6b, T8a, T8b), and potential alcohol is significantly increased by all of the treatments including late harvest, especially with a low number of buds and with both defoliation and bunch thinning performed (T4b).

The weight of wood generally increases due to a high number of buds. Conversely, a low number of buds entails a decrease in the weight of wood, which is stronger if treatment T4a is applied (low number of buds, both defoliation and bunch thinning performed, technological harvest).

The impact of treatments on SPAD measured in June is stronger in all treatments with a high number of buds: a greater number of sprouts, and consequently of leaves, delays the maturation of the leaves themselves, which in turn show a lower nitrogen and chlorophyll content in June (less intense green).

Total acidity and pH are mainly influenced by harvest time: late harvest produces less acid grapes (higher pH). An interesting exception is treatment T5b (high number of buds, defoliation and grape thinning not performed, late harvest), which has a higher total acidity than the reference treatment.

Potential alcohol is positively influenced by each treatment, excepting T5b (high number of buds, defoliation and bunch thinning not performed), with a stronger effect due to treatments including late harvest.

The anthocyanin content is worsened by a high number of buds. Treatment T4a (low number of buds, defoliation and grape thinning performed, technological harvest) is a very interesting option to emphasize the richness in color of wine.

The best treatments to obtain wines with good structure (high polyphenolic content) are those with a low number of buds (T1a, T1b, T2a, T2b, T3a, T3b, T4a, T4b). Another valid option for this objective is treatment T8b (high number

of buds, defoliation and grape thinning performed, late harvest). The best treatments to obtain fresh wines suitable for young consumers are those with technological rather than late harvest, preferably with a high number of buds, and anyway with both defoliation and grape thinning not performed.

#### 4. Discussion

The wine market demands frequent style adjustments. In some cases, “meditation” wines with high alcohol content, richness in color and great tannin structure are required. In other cases, “handy” wines (good for all occasions), characterized by moderate alcohol content, low astringency and polyphenolic content, are demanded. Although many varieties of wines can be obtained from different vinification procedures applied to the same grapes, several important features like alcohol content and yield per hectare cannot be achieved without relying on special agronomic settings. This issue implies relevant consequences for winery revenues: if the objective is to produce a simple wine, it is better to aim at a high yield per hectare, because such wine will be sold at a low price. Conversely, when aiming at producing a complex wine, the control of polyphenolic content is much more important, so a lower yield is accepted because the wine will have a higher value. On these grounds, new techniques to check whether the agronomic setting is achieving the objective before completing the grape harvest could be of great help for viticulturists.

In this paper, important insights towards this direction were provided by applying a Conditional Linear Gaussian Network (CLGN) model. On one hand, we evaluated the effectiveness of a set of canopy management techniques combined with two different harvest times on multiple outcomes representing the quality of grapes by exploiting conditional independence relationships learnt from data, resulting in a remarkable reduction in the number of parameters. On the other hand, the CLGN model made it possible to find the minimal set of predictors for each dimension, a very useful piece of information to establish the moment when the effectiveness of treatments on any single outcome can be

predicted. Interestingly, we found that the impact of the considered treatments on the anthocyanin content, and thus on the color of wine, can be predicted even in the vegetative season of the grapevine.

The values of R-squared indices are not high in several regression models. If we exclude the possibility that the investigated phenomenon has an inherently large unstructured variability, a natural explanation is that some informative explanatory variables have been omitted because they are unobserved. Omitted variables may entail bias in parameter estimates, but notably, the signs of the estimated parameters agree with tentative causal explanations, although orthogonality between observed and omitted covariates is not expected in general. Furthermore, expert prior information was exploited through a ‘blacklist’, to exclude from consideration all regression models breaking established causal knowledge, and through a ‘whitelist’, to force the inclusion of variables with a widely accepted explanatory role.

In our work, we implicitly considered parameters as fixed, given that only two years of data were available. Extension of the model to include a random factor representing different years of experimentation could be considered to improve the quality of predictions and to determine whether each covariate might interact with the year of experimentation in each regression model.

### **Acknowledgements**

This study was part of a research project funded by the Italian Ministry of Economic Development and Consorzio Tuscania (<http://www.ricercatuscania.it>), Florence, Italy. This study was partially supported by the University of Florence, funding framework *Progetto strategico di ricerca di base per l'anno 2015*, grant *Disegno e analisi di studi sperimentali e osservazionali per le decisioni in ambito epidemiologico, socio-economico, ambientale e tecnologico*.

## REFERENCES

- Kontoudakis N., Esterueals M., Fort F., Canalis J.M., Zamora F. (2011): Use of Unripe Grapes Harvested During Cluster Thinning as a Method for Reducing Alcohol Content and pH of Wine. *Australian Journal of Grape and Wine Research* 17(2): 230–238.
- Lauritzen S.L., Dawid A.P., Larsen B.N., Leimer H.G. (1990): Independence properties of directed Markov fields. *Networks* 20: 491–505.
- Lauritzen S.L., Wermuth N. (1989): Graphical models for associations between outcomes, some of which are qualitative and some quantitative. *The Annals of Statistics* 17: 31–57.
- Lauritzen S.L. (1999). *Graphical Models*. Oxford University Press, Oxford, UK.
- Pearl J. (2009): Causal inference in statistics: An overview. *Statistics Surveys* 3: 96-146.
- Ribereau-Gayon P., Glories Y., Maujean A., Dubourdieu D. (1999): *Handbook of Enology*, Vol. 2, John Wiley & Sons, Chichester, UK.
- R Core Team (2016): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, AT. ISBN 3-900051-07-0. <http://www.R-project.org/>
- Scutari M. (2010): Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software* 35(3): 1–22.
- Tian J., Paz A., Pearl J. (1998): Finding Minimal D-Separators. Technical Report R-254, UCLA Computer Science, Los Angeles, US-CA.
- Wright S. (1934): The method of path coefficients. *Annals of Mathematical Statistics*. 5(3): 161–215.

## APPENDIX.

**Summary of parameter estimation**

SproutN

|             | Estimate  | Std. Error | t value | Pr(> t ) |     |
|-------------|-----------|------------|---------|----------|-----|
| (Intercept) | -0.145467 | 0.027090   | -5.370  | 0.0000   | *** |
| T1b         | 0.011710  | 0.037207   | 0.315   | 0.7531   |     |
| T2a         | -0.002115 | 0.038563   | -0.055  | 0.9563   |     |
| T2b         | 0.012868  | 0.037621   | 0.342   | 0.7324   |     |
| T3a         | 0.020790  | 0.038312   | 0.543   | 0.5875   |     |
| T3b         | 0.047916  | 0.037207   | 1.288   | 0.1983   |     |
| T4a         | 0.031949  | 0.038071   | 0.839   | 0.4017   |     |
| T4b         | 0.072668  | 0.037207   | 1.953   | 0.0512   | .   |
| T5a         | 0.207992  | 0.038312   | 5.429   | 0.0000   | *** |
| T5b         | 0.279858  | 0.037410   | 7.481   | 0.0000   | *** |
| T6a         | 0.201257  | 0.038071   | 5.286   | 0.0000   | *** |
| T6b         | 0.261139  | 0.037207   | 7.018   | 0.0000   | *** |
| T7a         | 0.271160  | 0.038071   | 7.122   | 0.0000   | *** |

|     |          |          |       |        |     |
|-----|----------|----------|-------|--------|-----|
| T7b | 0.253153 | 0.037621 | 6.729 | 0.0000 | *** |
| T8a | 0.269138 | 0.038312 | 7.025 | 0.0000 | *** |
| T8b | 0.293167 | 0.037621 | 7.793 | 0.0000 | *** |

Residual standard error: 0.1692 on 643 degrees of freedom.  
R-squared: 0.3335

#### BunchN

|             | Estimate  | Std. Error | t value | Pr(> t ) |     |
|-------------|-----------|------------|---------|----------|-----|
| (Intercept) | 0.135612  | 0.054308   | 2.497   | 0.0128   | *   |
| T1b         | 0.034441  | 0.072976   | 0.472   | 0.6371   |     |
| T2a         | -0.223765 | 0.075629   | -2.959  | 0.0032   | **  |
| T2b         | -0.305478 | 0.073789   | -4.140  | 0.0000   | *** |
| T3a         | -0.019026 | 0.075153   | -0.253  | 0.8002   |     |
| T3b         | 0.038596  | 0.073065   | 0.528   | 0.5975   |     |
| T4a         | -0.362504 | 0.074706   | -4.852  | 0.0000   | *** |
| T4b         | -0.324205 | 0.073187   | -4.430  | 0.0000   | *** |
| T5a         | 0.102694  | 0.076839   | 1.336   | 0.1819   |     |
| T5b         | 0.072903  | 0.076494   | 0.953   | 0.3409   |     |
| T6a         | -0.279109 | 0.076270   | -3.659  | 0.0003   | *** |
| T6b         | -0.313844 | 0.075714   | -4.145  | 0.0000   | *** |
| T7a         | -0.002341 | 0.077555   | -0.030  | 0.9759   |     |
| T7b         | 0.033386  | 0.076336   | 0.437   | 0.6620   |     |
| T8a         | -0.325812 | 0.077966   | -4.179  | 0.0000   | *** |
| T8b         | -0.387063 | 0.077188   | -5.015  | 0.0000   | *** |
| SproutN     | 0.924296  | 0.077342   | 11.951  | 0.0000   | *** |

Residual standard error: 0.3318 on 642 degrees of freedom.  
R-squared: 0.3831

#### Grapew

|             | Estimate  | Std. Error | t value | Pr(> t ) |     |
|-------------|-----------|------------|---------|----------|-----|
| (Intercept) | 0.003785  | 0.011279   | 0.336   | 0.7373   |     |
| SproutN     | -0.233823 | 0.069638   | -3.358  | 0.0008   | *** |
| BunchN      | 0.779744  | 0.033113   | 23.548  | 0.0000   | *** |
| woodw       | 0.382944  | 0.037363   | 10.249  | 0.0000   | *** |
| NDVI06      | 0.819124  | 0.128271   | 6.386   | 0.0000   | *** |
| NDVI08      | 0.278226  | 0.105340   | 2.641   | 0.0085   | **  |
| Acid        | -1.099739 | 0.140918   | -7.804  | 0.0000   | *** |
| Brix        | -1.424193 | 0.192959   | -7.381  | 0.0000   | *** |
| pH          | -2.816734 | 0.577085   | -4.881  | 0.0000   | *** |
| Anthoc      | -0.138574 | 0.036960   | -3.749  | 0.0002   | *** |



Residual standard error: 0.2889 on 649 degrees of freedom.  
R-squared: 0.7957

woodw

|             | Estimate | Std. Error | t value | Pr(> t ) |     |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | -0.01059 | 0.01048    | -1.011  | 0.3125   |     |
| SproutN     | 0.20867  | 0.06351    | 3.286   | 0.0011   | **  |
| BunchN      | 0.10552  | 0.02820    | 3.742   | 0.0002   | *** |
| SPAD06      | 1.32072  | 0.13496    | 9.786   | 0.0000   | *** |
| SPAD08      | 1.17073  | 0.12261    | 9.549   | 0.0000   | *** |
| NDVI08      | 0.69730  | 0.09308    | 7.491   | 0.0000   | *** |

Residual standard error: 0.2683 on 653 degrees of freedom.  
R-squared: 0.6805

SPAD06

|             | Estimate  | Std. Error | t value | Pr(> t ) |     |
|-------------|-----------|------------|---------|----------|-----|
| (Intercept) | 0.075792  | 0.015201   | 4.986   | 0.0000   | *** |
| T1b         | -0.002311 | 0.020426   | -0.113  | 0.9100   |     |
| T2a         | -0.008399 | 0.021169   | -0.397  | 0.6917   |     |
| T2b         | 0.003380  | 0.020654   | 0.164   | 0.8700   |     |
| T3a         | -0.035988 | 0.021036   | -1.711  | 0.0876   | .   |
| T3b         | -0.032887 | 0.020451   | -1.608  | 0.1083   |     |
| T4a         | -0.050088 | 0.020910   | -2.395  | 0.0169   | *   |
| T4b         | -0.053681 | 0.020485   | -2.620  | 0.0090   | **  |
| T5a         | -0.115331 | 0.021507   | -5.362  | 0.0000   | *** |
| T5b         | -0.146345 | 0.021411   | -6.835  | 0.0000   | *** |
| T6a         | -0.096360 | 0.021348   | -4.514  | 0.0000   | *** |
| T6b         | -0.117209 | 0.021193   | -5.531  | 0.0000   | *** |
| T7a         | -0.156668 | 0.021708   | -7.217  | 0.0000   | *** |
| T7b         | -0.134990 | 0.021367   | -6.318  | 0.0000   | *** |
| T8a         | -0.126651 | 0.021823   | -5.804  | 0.0000   | *** |
| T8b         | -0.138303 | 0.021605   | -6.401  | 0.0000   | *** |
| SproutN     | 0.433518  | 0.021648   | 20.026  | 0.0000   | *** |

Residual standard error: 0.09287 on 642 degrees of freedom.  
R-squared: 0.3925

NDVI06

|             | Estimate | Std. Error | t value | Pr(> t ) |     |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 0.000906 | 0.003570   | 0.254   | 0.7998   |     |
| SproutN     | 0.052454 | 0.019907   | 2.635   | 0.0086   | **  |
| SPAD06      | 0.439185 | 0.034649   | 12.675  | 0.0000   | *** |

Residual standard error: 0.09162 on 656 degrees of freedom.  
R-squared: 0.2842

## SPAD08

|             | Estimate | Std. Error | t value | Pr(> t )   |
|-------------|----------|------------|---------|------------|
| (Intercept) | 0.003414 | 0.003391   | 1.007   | 0.3140     |
| SPAD06      | 0.657233 | 0.033891   | 19.392  | 0.0000 *** |
| NDVI06      | 0.347170 | 0.036889   | 9.411   | 0.0000 *** |

Residual standard error: 0.08702 on 656 degrees of freedom.  
R-squared: 0.5805

## NDVI08

|             | Estimate  | Std. Error | t value | Pr(> t )   |
|-------------|-----------|------------|---------|------------|
| (Intercept) | -0.004859 | 0.004378   | -1.110  | 0.2675     |
| SproutN     | 0.104257  | 0.023594   | 4.419   | 0.0000 *** |
| NDVI06      | 0.135986  | 0.050226   | 2.707   | 0.0070 **  |
| SPAD08      | 0.431695  | 0.041861   | 10.313  | 0.0000 *** |

Residual standard error: 0.1123 on 655 degrees of freedom.  
R-squared: 0.3319

## Acid

|             | Estimate   | Std. Error | t value | Pr(> t )   |
|-------------|------------|------------|---------|------------|
| (Intercept) | 0.0009448  | 0.0047161  | 0.200   | 0.8413     |
| SproutN     | -0.0859214 | 0.0286893  | -2.995  | 0.0028 **  |
| BunchN      | 0.0745025  | 0.0135083  | 5.515   | 0.0000 *** |
| SPAD06      | -0.3684176 | 0.0525611  | -7.009  | 0.0000 *** |
| NDVI06      | -0.1935515 | 0.0552518  | -3.503  | 0.0005 *** |
| NDVI08      | -0.2441154 | 0.0419743  | -5.816  | 0.0000 *** |
| Brix        | -0.6116201 | 0.0682666  | -8.959  | 0.0000 *** |
| Anthoc      | -0.0684740 | 0.0222445  | -3.078  | 0.0022 **  |
| Polyph      | 0.1803238  | 0.0302365  | 5.964   | 0.0000 *** |

Residual standard error: 0.1209 on 650 degrees of freedom.  
R-squared: 0.3793

## Potass

|             | Estimate  | Std. Error | t value | Pr(> t )   |
|-------------|-----------|------------|---------|------------|
| (Intercept) | -0.005046 | 0.005670   | -0.890  | 0.3738     |
| BunchN      | -0.071803 | 0.014636   | -4.906  | 0.0000 *** |

|        |          |          |       |        |     |
|--------|----------|----------|-------|--------|-----|
| SPAD06 | 0.391810 | 0.052731 | 7.430 | 0.0000 | *** |
| Anthoc | 0.061200 | 0.016229 | 3.771 | 0.0002 | *** |

Residual standard error: 0.1454 on 655 degrees of freedom.  
R-squared: 0.1052

#### Brix

|             | Estimate  | Std. Error | t value | Pr(> t ) |     |
|-------------|-----------|------------|---------|----------|-----|
| (Intercept) | -0.063000 | 0.009919   | -6.352  | 0.0000   | *** |
| T1b         | 0.083582  | 0.013632   | 6.132   | 0.0000   | *** |
| T2a         | 0.024226  | 0.014115   | 1.716   | 0.0866   | .   |
| T2b         | 0.095053  | 0.013820   | 6.878   | 0.0000   | *** |
| T3a         | 0.048665  | 0.014022   | 3.471   | 0.0005   | *** |
| T3b         | 0.122648  | 0.013620   | 9.005   | 0.0000   | *** |
| T4a         | 0.044474  | 0.013934   | 3.192   | 0.0015   | **  |
| T4b         | 0.122709  | 0.013616   | 9.012   | 0.0000   | *** |
| T5a         | -0.005952 | 0.014023   | -0.424  | 0.6714   |     |
| T5b         | 0.058604  | 0.013704   | 4.276   | 0.0000   | *** |
| T6a         | 0.023181  | 0.013938   | 1.663   | 0.0968   | .   |
| T6b         | 0.083249  | 0.013650   | 6.099   | 0.0000   | *** |
| T7a         | 0.030893  | 0.013936   | 2.217   | 0.0270   | *   |
| T7b         | 0.097088  | 0.013768   | 7.051   | 0.0000   | *** |
| T8a         | 0.062341  | 0.014029   | 4.444   | 0.0000   | *** |
| T8b         | 0.121227  | 0.013768   | 8.805   | 0.0000   | *** |
| Anthoc      | 0.087758  | 0.006323   | 13.880  | 0.0000   | *** |

Residual standard error: 0.06191 on 642 degrees of freedom.  
R-squared: 0.4204

#### pH

|             | Estimate   | Std. Error | t value | Pr(> t ) |     |
|-------------|------------|------------|---------|----------|-----|
| (Intercept) | -0.0004402 | 0.0006560  | -0.671  | 0.5024   |     |
| SproutN     | 0.0115191  | 0.0037932  | 3.037   | 0.0025   | **  |
| Woodw       | 0.0061100  | 0.0023061  | 2.650   | 0.0083   | **  |
| SPAD06      | 0.0407614  | 0.0086633  | 4.705   | 0.0000   | *** |
| Acid        | -0.1814311 | 0.0051612  | -35.153 | 0.0000   | *** |
| Potass      | 0.0569803  | 0.0044584  | 12.780  | 0.0000   | *** |
| Brix        | 0.1606425  | 0.0097056  | 16.552  | 0.0000   | *** |
| Anthoc      | -0.0239600 | 0.0030594  | -7.832  | 0.0000   | *** |
| Polyph      | 0.0210688  | 0.0041708  | 5.051   | 0.0000   | *** |

Residual standard error: 0.01681 on 650 degrees of freedom.  
R-squared: 0.8389

## Anthoc

|             | Estimate  | Std. Error | t value | Pr(> t ) |     |
|-------------|-----------|------------|---------|----------|-----|
| (Intercept) | 0.006827  | 0.012385   | 0.551   | 0.5820   |     |
| BunchN      | -0.136132 | 0.031878   | -4.270  | 0.0000   | *** |
| woodw       | -0.330309 | 0.033425   | -9.882  | 0.0000   | *** |
| NDVI08      | -0.452975 | 0.111038   | -4.079  | 0.0000   | *** |

Residual standard error: 0.3177 on 655 degrees of freedom.  
R-squared: 0.3364

## Polyph

|             | Estimate  | Std. Error | t value | Pr(> t ) |     |
|-------------|-----------|------------|---------|----------|-----|
| (Intercept) | 0.002074  | 0.006110   | 0.340   | 0.7343   |     |
| BunchN      | 0.055373  | 0.016173   | 3.424   | 0.0006   | *** |
| NDVI06      | -0.373442 | 0.063648   | -5.867  | 0.0000   | *** |
| NDVI08      | 0.234360  | 0.052270   | 4.484   | 0.0000   | *** |
| Brix        | 0.286761  | 0.087584   | 3.274   | 0.0011   | **  |
| Anthoc      | 0.550065  | 0.018783   | 29.285  | 0.0000   | *** |

Residual standard error: 0.1566 on 653 degrees of freedom.  
R-squared: 0.6658