Longitudinal Analysis of symptom expression in grapevines affected by esca

Federico Mattia Stefanini[1], Giuseppe Surico[2] and Guido Marchi[2]

[1]Dipartimento di Statistica "G. Parenti", Università, Viale Morgagni 59, 50134 Firenze, Italy
[2]Dipartimento di Biotecnologie Agrarie - Patologia vegetale, Università,
Piazzale delle Cascine, 28, 50144 Firenze, Italy

Summary. An analysis of symptom expression in esca infected grapevines was performed by focusing on the dynamics of each plant. A parametric statistical model was proposed to evaluate the probability that a plant would show esca symptoms at given values for a relevant set of factors (year, presence of symptoms in the previous year, presence of plants with symptoms in the close neighborhood). The statistical tests of the hypotheses revealed that the considered factors explained a large amount of the observed variability. In particular, the state of plants in the close vicinity is one of those factors. Thus we found evidence that there was an association between plant vicinity and esca symptoms. Future developments of our model will include the factors field column and weather.

Key words: esca, grapevine, correlated data, generalized linear model.

Introduction

Esca is a complex disease. At least 2 or 3 fungi (*Phaeoacremonium chlamydosporum, P. aleophilum* and *Fomitiporia punctata*), acting in succession or in combination, are involved. Each fungus produces in the wood of the trunk different symptoms and a multitude of interactions can occur due to various combinations of fungus or fungi, host and environment. One of the results of these interactions is the characteristic discontinuity in symptom expression of esca.

Esca can increase rapidly in the vineyard so that almost all the vines may be infected by the time the vineyard reaches 25-30 years of age. The mode of spread of the causal agents of the disease differs. *F. punctata* is not believed to be transmitted from vine to vine (Cortesi *et al.*, 2000). On the other hand, it seems that *Phaeoacremonium* spp. may spread in the vineyard from external or internal sources of inoculum (Larignon, 1999).

Spatial studies of esca have been conducted previously to quantify the spatial aspects of the disease in regular lattices of host vines. Due to the discontinuity in symptom expression the analysis of patterns was in practice restricted to data cumulated over several years of observations (Surico *et al.*, this issue). This meant that hypotheses concerning the details of disease increase over time could not be tested directly.

The quantitative analysis of symptom expression over time requires parametric statistical models, especially if inferential statements are needed besides descriptive summaries. A next-year forecast of symptom incidence must be based

Corresponding author: F.M. Stefanini
Fax: +39 055 4223 560
E-mail: stefanin@ds.unifi.it

on the point estimate of meaningful parameters, here represented by transition probabilities from the field state of one year to the field state of the next (presence/absence of symptoms or death). Moreover, model selection involves the identification of risk factors and protective factors, i.e. of sources of variation associated with a modification of the baseline probability of displaying esca symptoms.

In this paper we present a longitudinal model of symptom expression in which the probability of showing esca symptoms changes according to field factors. Model selection was based on the Akaike Information Criterion (AIC) (McCullagh and Nelder, 1989), and statistical tests of hypotheses were performed looking at differences of deviance values. The results of the analysis suggested that the presence of diseased plants in the close (column) neighborhood of a plant one year was associated with a change of the probability that that plant would show esca symptoms the next year. The model detailed in the next paragraphs refers to disease symptoms, but further improvements, now in progress, will deal with the latent variable that accounts for sickness-health-death besides symptom expression.

Materials and methods

The proposed longitudinal model was applied to a vineyard (GTFI) in the province of Florence. During the summer of each year from 1993 to 1998 a block of vines, 10 columns x 51-62 rows, was surveyed in the vineyard. The collection of data and the characteristics of GTFI are described in Surico *et al.* (this issue). A dataset was obtained by coding survey results in three classes, Y=0 for absence of symptoms, Y=1 for presence of symptoms and Y=2 for dead plants.

Plant symptoms for years 2 to 6, i.e. 5 yearly transitions, were included in the model as dependent variables.

Model factors were coded using dummy variables (auxiliary variables). The neighborhood of a plant consisted of one plant on either side along the column. The variable DD (plant pairs) had value 1 if the plant preceding and/or following a given plant showed symptoms, 0 if it did not. Dead plants were ignored and when this happens the neighborhood of the living plants was extended to include the next living plant. The variables YJ and Y (state of the plant) have the same interpretation but YJ refers to the value Y lagged back one year. The variable CO (column) has the values 1,2, ..., 10, according to the column number in the field. The variable TRA (transition) has values 1,2,…,5 respectively for the transitions first-to-second year, second-to-third year, and so on.

Statistical analysis

We used a longitudinal model in which the probability that a plant would show esca symptoms in a year t depended on the field state in year t-1 (t=2,3,…,6, transition models) (Diggle *et al.*, 1994). The field state was determined by the value of those field factors described in the paragraph above, that is by YJ, DD, TRA and CO.

In the simplest model, the probability $P[\mathrm{Y=i}]=\pi_i$ that a plant belongs to the class coded as Y= i – 1 in a given year does not depend on the value of the field factors in the previous year. In this case the vector of parameters (π_1, π_2, π_3), with $\pi_1+\pi_2+\pi_3=1$, fully specifies the model. The simplest model is certainly wrong, because dead plants will remain dead the next year, and therefore the vector above does not suffice. Another, enlarged model has six parameters, three that represent probability values if YJ=0 (no symptoms in the previous year) and three parameters if YJ=1 (symptoms in the previous year). It is clear that YJ=2 implies Y=2, that is, the last three probability values (parameters) are known to be 0, 0 and 1. In that case the matrix of transition probabilities will be

$$\begin{bmatrix} \pi_{11} & \pi_{12} & \pi_{13} \\ \pi_{21} & \pi_{22} & \pi_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

where in the first row YJ=0, in the second YJ=1 and in the third YJ=2. On columns (left to right), the values are Y=0, Y=1 and Y=2. Note that probability values on each row add up to one, $\sum_j \pi_{i,j} = 1$ for i=1, 2. Moreover, a compact notation for the matrix above does not contain the last row, because it consists of numbers rather than unknown parameters.

The probability transition matrix of the enlarged model does not depend on years (transitions, TRA) or on the state of the plant neighborhood (DD); therefore the same matrix is assumed to hold in different years, and in general, for different values of field factors. If this assumption is wrong the matrix must be enlarged adding more rows to allow for different transition probabilities under different field conditions (combinations of factors YJ, DD, TRA). In other words, a model must be formulated to define the features of the probability transition matrix (number and features of each row).

Generalized linear models are linear models on a transformed scale (Mc Cullagh and Nelder, 1989). We simplify the notation by always using π_1, π_2, π_3 to denote the probability that a plant will belong to class 0, 1 or 2, and by explicitly stating in each case the value of the conditioning variables (DD, TRA, etc.), that is, the row of the probability transition matrix involved in a formula.

In the logistic multinomial model, the probability **P[**Y=i**]**=π_i that a plant belongs to the class coded as Y=i–1 is transformed in the product of π_1 times exp(η), that is $\pi_i = \pi_1 \cdot \exp(\eta)$. Here, π_1 is the probability associated with the class Y=0 that is taken as reference (baseline probability). Boundaries in the range of probability values are removed from the logarithmic scale because log $(\pi_i/\pi_1) = \eta$ has rank (– ∞, ∞).

The linear predictors η may be specified as constants (e.g. η_1=0, η_2 and η_3 for Y=0, Y=1 and Y=2 respectively), or as the sum of several parameters. In the latter case, η contains a linear combination of effects associated with field factors and may also include interaction terms. For example, let α_j be the effect for TRA and γ_m the effect for DD, then the saturated linear model for these two factors is $\alpha_j + \gamma_m + \alpha\gamma_{jm}$, where α_j is the effect for the year transition labelled as j, γ_m the neighborhood factor with level m, and $\alpha\gamma_{jm}$ the interaction term.

Note that the probability values π_i, i=1,2,3 add to one, that is $\pi_1+\pi_2+\pi_3=1$. From this last equation p_1 can be defined as a function of linear predictors, that is, $\pi_1 = \dfrac{1}{\sum_i \exp(\eta_i)}$.

Therefore, the probability associated with the factor level i-1 of Y is also defined by:

$$\pi_i = \frac{\exp(\eta_i)}{\sum_r \exp(\eta_r)}$$

Model parameters may be interpreted on the multiplicative scale, and their values are obtained by exponentiation, e.g.:

$$\exp(\eta_i) = \exp\left(\alpha_j + \gamma_m + \alpha\gamma_{jm}\right) = \\ = \exp(\alpha_j) * \exp(\gamma_m) * \exp(\alpha\gamma_{jm})$$

Under the introduced “treatment” parameterization exp(α_2) is the amount of change due to a shift from the base-line class with TRA=1, to the class TRA=2. This choice of parameterization implies α_1=0 (parameter constraint), and it also holds for the first level of the other model factors (and interactions).

Model fitting was performed using a stepwise search in the model space to minimize the value of the Akaike’s Information Criterion -2*log-likelihood + 2*npar. The AIC value of model M was small if the log-likelihood value of M was large and if the number of parameters contained in M (npar) was small, so the higher the number of parameters, the larger the penalization term. Plots of deviance and Pearson residuals were inspected to find patterns indicating violation of the model assumptions.

Results and discussion

The stepwise model selection procedure ended in the model Y ~ TRA + YJ+ DD + YJ:DD (Wilkinson and Rogers notation in McCullagh and Nelder, 1989), where YJ:DD indicates the interaction of the two specified model terms: M1, with interaction; M2 without interaction.

The analysis of deviance for these two models is shown in Table 1. The null hypothesis “parameter equal to zero” was rejected for all but the last model term with significance level 5%. The AIC values for the best model were 96.8982, and it included the interaction YJ:DD. Nevertheless, the model without the interaction term had an AIC value of 98.7862. As expected, the value of residual deviance is close to the value of its degree of freedom, with or without inclusion of the interaction term.

We prefer the simpler model without interaction for several reasons. The interaction term is not significantly different from zero (α = 0.05), although the critical value is defined using asymptotic distributions. The AIC values are very similar, thus the decision to remove the interaction term from the model is sound. Moreover, the model interpretation is simpler without interaction terms,

Table 1. Analysis of deviance. Terms are added sequentially (first to last). The Table shows, from left to right: name of the model term, its degree of freedom (DF), its deviance (Dev.), the degree of freedom (DF) of the residual deviance, its value (Res. Dev.), and the probability associated to the observed test statistic. Finally, the AIC value for the two models used is also shown.

Model term	DF	Dev.	DF	Res. Dev.	Prob.	AIC
TRA	8	20.04	26	270.68	0.0102	
YJ	2	233.50	24	37.18	0.0000	
DD	2	6.39	22	30.78	0.0409	98.7862
YJ:DD	2	5.89	20	24.90	0.0526	96.8982

and in this case there did not seem to be any reasonable explanation for the interaction between YJ and DD from the standpoint of plant pathology. The estimated values of the interaction terms are both negative (result not shown): this is as if plants with symptoms in the neighborhood of a given plant prevent it from showing symptoms or death the following year if the plant already showed symptoms in the current year. Of course this does not make any sense from the point of view of phytopathology. In any case, another explanation might be looked for in the presence of diversity among plants (a latent variable). All in all, the exclusion of interaction terms is reasonable, but further research on this issue is needed.

Table 2. Values of the estimated parameters. The parameter name, the estimated value and the standard error (SE) are shown from left to right for the model without interaction.

Parameters	Value	SE
Y1	-2.0178	0.1312
Y2	-5.0189	0.4143
TRA2:Y1	-0.4504	0.1876
TRA3:Y1	0.2372	0.1701
TRA4:Y1	-0.0118	0.1755
TRA5:Y1	0.0007	0.1761
TRA2:Y2	0.1924	0.4367
TRA3:Y2	0.1268	0.5281
TRA4:Y2	0.8561	0.4023
TRA5:Y2	-0.2159	0.4989
YJY1	1.4078	0.1318
YJY2	3.4273	0.3196
DD:Y1	0.2361	0.1238
DD:Y2	-0.4494	0.3227

More insights on the two models are obtained by inspecting differences between expected and observed cell frequencies with and without interaction terms (Fig. 1). Generally, the differences are quite small, but for cells with small counts the difference is larger: only for a few of them the model with interaction is certainly better. The interaction term seems to be mostly related to an adjustment of small counts and of one atypical cell, so that its role could be due merely to sampling variability (noise), as indeed the statistical test suggests.

Note that we removed from the dataset those cells containing sampling zeros and those that are known in advance because of forbidden transitions (constrained parameters), e.g. dead plants which will not become alive in a following year.

The estimate of parameters under the model without interaction (M2) and their standard errors are shown in Table 2. The probability of showing disease symptoms changes with the year, the presence of symptoms a year earlier, and the state of the plants in the neighborhood. The magnitude of the change is maximal for the model factor YJ, as expected from the standpoint of plant pathology.

The graphical analysis of residuals (McCullagh and Nelder, 1989) showed moderate evidence for possible departures from the model assumptions (Fig. 2). Only one residual was quite large with the model without the interaction term.

We also investigated the field column as a model factor, but the best model we obtained had an AIC value equal to 509 (data not shown), that is, about 5 times as much as the best model without such factor. Moreover, 20 parameters were not estimated due to the presence of singularities. In oth-

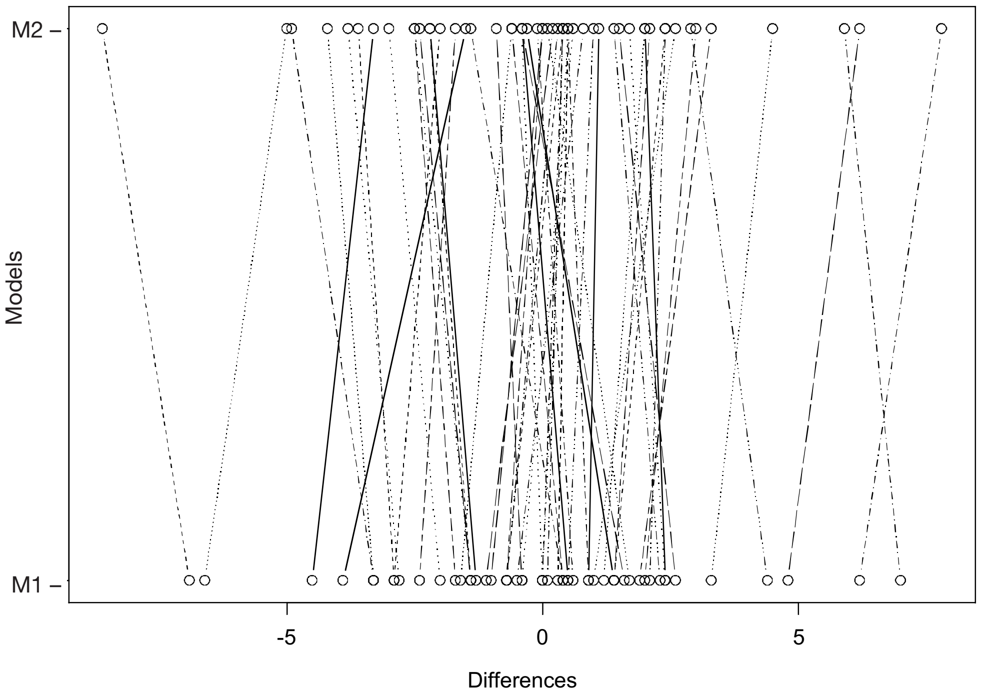

Fig. 1. Differences between models. Differences between expected and observed cell frequencies for model with (M1) and without (M2) interaction are shown in the plot. Straight lines connect values referred to the same cell under the two models.

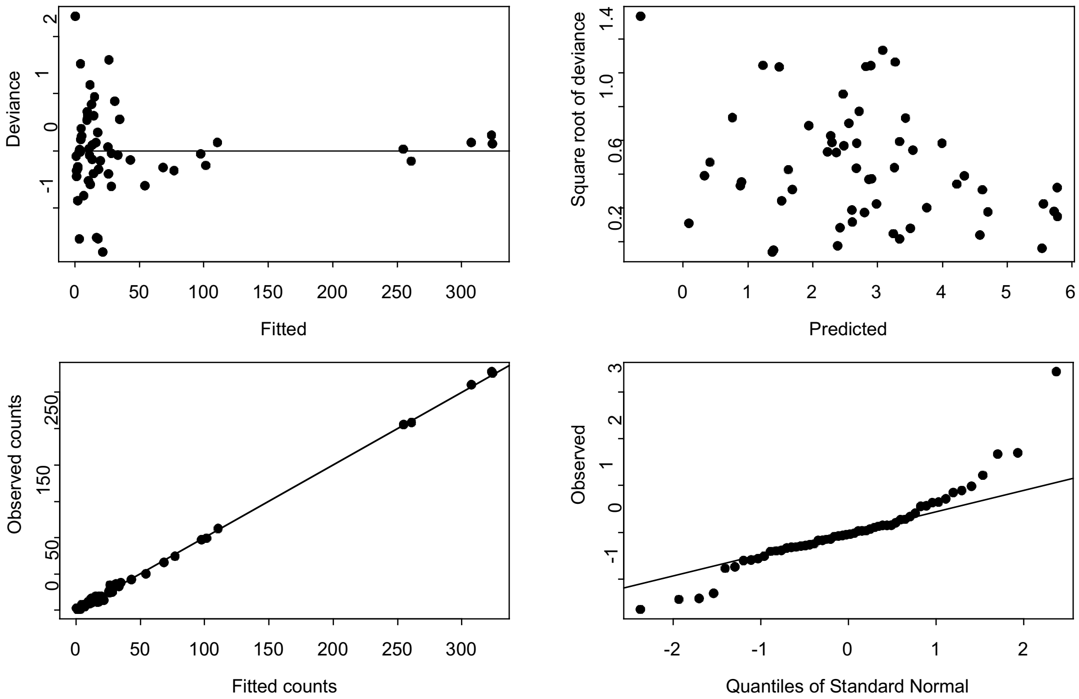

Fig. 2. Analysis of residuals. From top left to bottom right, several types of residuals are plotted: difference of deviance values specific for each cell against fitted values, a transformation of them (square root of the absolute value) vs. predicted values, observed against fitted frequencies, quantile-quantile plot of Pearson residuals.

Table 3. Conditional probability of symptom expression. The first 3 columns on the right show the field-state of a plant in a given year and the last 3 the probability of observing the value y of Y in the next year. The symbol NA is used to indicate cells that showed sampling zeros. Transitions from YJ = 2 to Y were removed from the Table because no transition is possible for dead plants.

TRA	YJ	DD	Y=0	Y=1	Y=2
1	0	0	0.8775	0.1167	0.0058
1	0	1	0.8559	0.1441	NA
1	1	0	0.5724	0.3110	0.1165
1	1	1	0.5501	0.3784	0.0714
2	0	0	0.9151	0.0775	0.0073
2	0	1	0.8989	0.0964	0.0046
2	1	0	0.6277	0.2174	0.1549
2	1	1	0.6266	0.2747	0.0987
3	0	0	0.8503	0.1433	0.0064
3	0	1	0.8241	0.1759	NA
3	1	0	0.5209	0.3587	0.1204
3	1	1	0.5341	0.4658	NA
4	0	0	0.8719	0.1145	0.0136
4	0	1	0.8501	0.1414	0.0084
4	1	0	0.4960	0.2663	0.2377
4	1	1	0.5036	0.3424	0.1540
5	0	0	0.8784	0.1169	0.0047
5	0	1	0.8558	0.1442	NA
5	1	0	0.5856	0.3184	0.0961
5	1	1	0.5577	0.3839	0.0584

er words, the number of cells in this model seemed too high for the available data. Moreover, many cell counts were smaller than 3 and were even zero. We did not investigate this model any further. The model output includes the evaluation of the conditional probability of observing the value y of Y given the current field state (Table 3). This Table is to be explained as a probability transition matrix. In the first four rows from the top the value of the field state during the first year (TRA=1) is shown: the plant has already shown symptoms before (YJ=1) or not (YJ=0), and the neighboring plant or plants carry (DD=1) or do not carry (DD=0) symptoms. On columns, the probabilities are: symptom expression (Y=1), death (Y=2) or no symptoms (Y=0). For example, the probability that in the second year a plant that did not show symptoms before (YJ=0) will show symptoms (Y=1) is 11,67% if DD=0 and 14.41% if DD=1. Thus, symptoms are more likely in a plant in which neighboring plants have shown symptoms. Similar considerations hold if the plant under consideration has already shown symptoms (YJ=1). Again, the probability that in the second year a symptomatic plant (YJ=1) will die (Y=2) is 11.65% if DD=0, and 7.14% if DD=1. This finding is somewhat in contrast with the preceding. In any case, as the results of Table 3 show, the probability that a symptomless plant (YJ=0) will not show any symptoms the following year (Y=0) is very high for all transitions (about 85-91% for DD=0, and about 82-89% for DD=1), and the probability is likewise high, though not as high as the first, that a symptomatic plant (YJ=1) will not show symptoms the following year (about 49-62% for DD=0 and about 50-62% for DD=1). Much lower on the other hand is the probability that a plant, whether symptomatic or not in one year, will show symptoms the following year: this occurs an average of 29% for DD=0 and 37% for DD=1 if the plant was symptomatic the year before, and about 11% for DD=0 and 14% for DD=1 if the plant had not been symptomatic. It is therefore much more probable that an asymptomatic plant in one year will also be asymptomatic a year later, than that a plant, symptomatic or not in one year, will be symptomatic the following year. Moreover, probability values were generally higher for DD=1. Lastly, the probability that an asymptomatic plant will die in a following year varies from 0.47% to 1.36% for DD=0 and from 0.46% to 0.84% (only 2 cases noted) for DD=1. For a symptomatic plant on the other hand such a probability increases to 23.77% (TRA=4) for DD=0 (average 14.51% for all transitions), and to 15.4% (TRA=4) for DD=1 (9.56% for all transitions). Interesting is further the fact that at the transition from year 2 to year 3 in all situations (DD=0; YJ=0, DD=1; YJ=1, DD=0; YJ=1, DD=1) the probability values were lower than the general averages. In fact, esca incidence was 15.96% in 1993, 16.82% in 1994 and only 11.11% in 1995 (see Table 3 in Surico *et al,*. this issue). Therefore the pattern of the probability values at different TRA values corresponded to variations in disease incidence over the years. This finding thus tallies with the survey results on esca incidence found in the vineyard, but others, discussed above, suggest for example that variations in the occurrence of symptoms are very probably dependent upon factors external to the plant: this is because an asymptomatic plant tends strongly (generally with more than 85% probability) to remain asymptomatic also the year following, while a symptomatic plant is much less likely (probability generally not more than 46%) to remain symptomatic a year later. Nevertheless, other inferences that can be based on the results shown in Table 3, such as that the health status of a plant (with or without symptoms) varies depending on whether neighboring plants are with or without symptoms, are not easy to explain from a plant pathological point of view.

Two more comments are required here. First, the statistical tests we used are based on asymptotic results and the number of plants is small in some cells: it was not possible to estimate the parameter in some cases. Thus further investigations are required to clarify the role of the interaction term after collecting more data. Second, it was not possible to include the field factor column (CO), thus some departures from the described probability transition matrix are expected after being able to handle an enlarged model that includes CO.

The model we have presented might be improved by introducing the latent variable "state of the plant" whose values should be "dead", "sick" or "healthy", not considered here, and weather parameters. Moreover, other definitions of neighborhood might be studied to test other hypotheses about the dynamics of symptom expression. Longer observation times should allow the study of field columns as a model factor. Random differences in soil features or differences induced by diversity of plants located in different columns might also be investigated.

Acknowledgements

The authors thank A. Gottard for valuable discussions on an early draft of the manuscript.

Literature cited

Cortesi P., M. Fischer and M.G. Milgroom, 2000. Population diversity of *Fomitiporia punctata* from grapevine and spread of esca disease. *In*: IOBC/wprs Bullettin, Working Group "Integrated Control in Viticulture", 1-4 March, Florence, Italy, (in press).

Diggle P.J., K. Liang and S.L. Zeger, 1994. *Analysis of Longitudinal data*. Clarendon Press, Oxford, UK.

Larignon P., 1999. Preliminary results on the biology of *Phaeoacremonium*. *In*: Black goo - Occurrence and Symptoms of Grapevine Declines. IAS/ICGTD Proceedings 1998 (L. Morton ed.), International Ampelography Society, Fort Valley, VA, USA, 49-55.

McCullagh P. and J.A. Nelder, 1989. *Generalized Linear Models*. Chapman & Hall, New York, USA.