Haulm senescence in potatoes and semi-parametric survival models

Sabine K. Schnabel[1,3], Paul H.C. Eilers[1,2],Paula Hurtado López[1,4,5], Richard G.F. Visser[3,4] and Fred A. van Eeuwijk[1,3]

[1] Biometris, Wageningen UR, Postbus 100, 6700 AC Wageningen, The Netherlands; sabine.schnabel@wur.nl (communicating author)

[2] Erasmus MC, Department of Biostatistics, Postbus 2040, 3000 CA Rotterdam, The Netherlands

[3] Center for Biosystems Genomics, Postbus 98, 6700 AB Wageningen, The Netherlands

[4] Wageningen UR, Plant Breeding, Postbus 386, 6700 AJ Wageningen, The Netherlands

[5] C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC), Wageningen UR, Droevendaalsesteeg 4, 6708 PB Wageningen, The Netherlands

Abstract: Haulm senescence describes the decay of potato plants on a discrete visual scale. A smooth semi-parametric hazard model is developed and applied to data from an agricultural experiment. Characteristics of the estimated hazard and survival function can be used to detect quantitative trait loci on chromosomes that can be associated with the senescence process.

Keywords: Hazard, P-splines, Poisson distribution, senescence QTL analysis

1 Introduction

An important phase in the life of a (commercial) potato plant is the decay of the haulm, the part above the ground. Only after complete decay potatoes will be harvested. In agricultural experiments, the state of the haulm, called its senescence, is monitored regularly and visually graded on a discrete scale. Breeders are interested in characterizing haulm senescence and in finding genes that influence it.

Traditionally senescence data have been modeled by parametric curves (Malosetti et al., 2006). In addition we have been working on semi-parametric alternatives (Hurtado at al., 2010). They increase the flexibility of the statistical model, but treat the data as independent observations of a time series. However, senescence is an irreversible process, so this model is lacking a strong biological basis. Here we introduce a semi-parametric survival model that is more realistic.

To improve the yield of potato plant breeders are interested in identifying the genomic regions underlying yield and yield-related traits. Such a region is called a quantitative trait locus (QTL). Plants that senesce more slowly retain higher photosynthesic capacity and thereby may reach higher yields. Statistical descriptions of the senescence process become useful for plant breeding purposes when they deliver process characterizations that are clearly and consistently different between potato plants with different genetic constitutions (genotypes). Identification of the genomic regions underlying senescence differences, so called QTL mapping, can be understood as the comparison between groups of genotypes with different DNA composition at specified positions in the genome for particular response traits such as yield and senescence characterizations. The statistical modeling and characterization of the senescence process is more successful when more and stronger QTLs are detected. For an example data set we show that our semi-parametric survival analysis approach was able to produce senescence characterizations for which clear QTLs turned up.

2 Data description

Our senescence data were obtained from a Finnish field experiment in 2004 with around 200 genotypes of a diploid potato population (Zaban et al., 2006). The senescence process of the haulms was recorded at different time points during the growing season. The status of the haulm y_i was scored on a scale from “green plant” ($y_i = 0$), “upper leaves with first signs of yellowing” ($y_i = 1$) etc. to “dead plant” ($y_i = 7$) (Celis-Gamboa et al., 2003).

To translate the senescence data to a survival model, we introduce an urn with seven marbles. At each step on the senescence scale a marble is removed. If monitoring had been done daily and long enough, we would know the times at which each change of state (removing of a marble) occurred. Then we would have had discrete survival data. In reality observations were made intermittently, approximately every 5 days (counted in days after planting: DAP). So the data are interval censored. We presently handle this complication by simply assuming the observed changes (zero, one, or more than one marble removed) to have occurred uniformly distributed over the respective time interval. We will return to this in the Discussion.

For each genotype three replicated plants were monitored. To simplify the analysis, their results were combined into one hypothetical urn with three times seven marbles.

The genotypes in the field experiment show a lot of variation in their behavior in terms of senescence as well as in other observed traits of the plants. Some genotypes develop early and therefore reach the state of “dead plant” (equivalent to zero marbles in the urn), while others develop late and might not reach the last state of the senescence scale. This diversity can also be seen in the final results.

3 Theory and Application

The time axis is divided into narrow intervals. In interval j we have the number at risk r_j and the number of events d_j. We model the logarithm of the hazard:

$$\log h_j = \sum_k b_{jk} \alpha_k, \tag{1}$$

where $B = [b_{jk}]$ is the matrix of $B-$spline basis functions and α the coefficient vector. The log-likelihood is

$$\log L = \sum_j d_j \log \mu_j - \sum_j \mu_j \tag{2}$$

with $\mu_j = \mathbf{E}(d_j) = r_j h_j$ the expected value of the number of events in interval j. $P-$splines include a penalty $\lambda ||D_3 \alpha||^2$ to get a smooth curve, where D_3 is a matrix that forms third order differences and λ a parameter to tune smoothness. After linearization we find that we have to solve the following system repeatedly:

$$(B^T \tilde{M} B + \lambda D_3^T D_3)\alpha = B^T(d - \tilde{\mu} + \tilde{M} B \tilde{\alpha}). \tag{3}$$

where $\tilde{M} = \text{diag}(\tilde{\mu})$. A more detailed description can be found in (Eilers, 1998).

In our application to haulm senescence of potato we estimate an individual hazard curve for each genotype in the population. Five examples of estimated smooth hazard curves along with their respective fitted and empirical survival curves are shown in Figure 1. As mentioned above we can see quite a range of different shapes for the hazard curves as well as for the survival curves. While the first two genotypes (CE140 and CE691) are commonly classified as intermediate, the other genotypes are developing early. However, we see that in this group the shape of hazard curves varies substantially. While CE102 and CE155 show a unimodal shape, CE685 tends to a bimodal hazard. In this case the senescence process seems to level off after an initial period of aging. The hazard increases then again towards the end of the observation period.

3.1 Analysis of quantitative trait loci

In order to identify possible QTLs influencing haulm senescence we can use the fitted curves. To characterize the curves we determine the mode of the hazard and the time point when it occurs as well as the time point when 1/5 of the senescence process is over (indicated by the horizontal lines in the survival curves of Figure 1). With these characteristics we performed a non-parametric QTL analysis using the rank sum test of Kruskal-Wallis available in MapQTL 6 (van Ooijen, 2009). QTL mapping was performed separately on the maternal and paternal maps (C and E respectively) and the criterion for detecting QTLs was set by a significance level of $p \leq 0.005$. We detected a major QTL on chromosome 5 related with the mode of the hazard, a QTL on chromosome 4 of the C parent related with the time point when 1/5 of the senescence process occurred as well as minor QTLs on other chromosomes. These findings are in line with previous research (Hurtado et al., 2010). The QTL on chromosome 4 related with the time point when 1/5 of the senescence process elapsed leads to an interpretation of that time point as the onset of senescence.

4 Discussion

To our knowledge, survival models have not been used for haulm senescence. We believe they offer a more realistic basis for the analysis of development traits over time than a curve fitting approach whether parametric (Malosetti et al., 2006) or semi-parametric (Hurtado at al., 2010). On the other hand we are aware that we only have scratched the surface of this field. Here we shortly discuss in which directions our research will be extended.

In this analysis we treated interval censoring in a very simplified manner and assumed a uniform distribution of events inside a time interval. In principle it is straightforward to write down a model in which a smooth hazard *and* a smoothly changing risk set (derived from that hazard) define the likelihood.

All marbles in the urn were considered exchangeable. In reality we have a multi-state model with a chain of states. The relative hazards of the changes of states most probably are not equal. By combining data from different genotypes in the same experiment, or from the same genotype in different experiments, or both, it might be possible to estimate an inflation or deflation factor for each state that is modulated by the overall smooth hazard.

Haulm senescence is strongly influenced by temperature. A real challenge will be to develop models in which genotype-specific parameters and observed temperature are combined. This will involve large-scale mixed model technology to combine data from different genotypes and different environments. Generalized additive models for the log-hazard, with additional shrinkage, might be a promising starting point.

Furthermore we also want to explore transformations of the time axis. The conventional choice of “days after planting” as time scale does not take into account temperature or photo period. We are currently experimenting with time measures including these factors such as thermal time and extensions to it. Thermal time is a measure for daily accumulations of heat taking into account the minimum and maximum temperature for growth of the particular species. This cumulative measure for heat can be further

extended to include the daylight length which is another important factor in the development of the plant.
We will also need good summaries of the results from our semi-parametric survival model to serve the needs of potato breeders well. In the present model the height, the position and the width of the hazard curve are candidates, but in more complex model the choice of characterizations may be less obvious.

References

Celis-Gamboa, C., Struik, P.C., Jacobsen, E., and Visser, R.G.F. (2003). Temporal dynamics of tuber formation and related processes in a crossing population of potato (Solanum tuberosum). *Annals of Applied Biology*, **143**, 175-186.

Eilers, P.H.C. (1998). Hazard smoothing with B–splines. In: Statistical Modeling. Proceedings of the 13th International Workshop on Statistical Modelling. 200-207, New Orleans, USA.

Hurtado, P., Schnabel, S., Zaban, A., Veteläinen, M., Virtainen, E. , Eilers, P., van Eeuwijk, F., Visser, R., and Maliepaard, C. (2010). Dynamics of senescence-related QTL in potato. *Under review.*

Malosetti, M., Visser, R.G.F., Celis-Gamboa, C., and van Eeuwijk, F.A. (2006). QTL methodology for response curves on the basis of non-linear mixed models, with an illustration to senescence in potato. *Theoretical Applied Genetics*, **113**, 288-300.

van Ooijen, J.W. (2009). MapQTL 6, Software for the mapping of quantitative trait loci in experimental populations of diploid species. Kyazma B.V., Wageningen, Netherlands.

Zaban, A., Veteläinen, M., Celis-Gamboa, C.B., van Berloo, R., Häggman, H., Visser, R.G.F. (2006). Physiological and genetical aspects of the broad based potato population (Solanum tuberosum L.) in the Netherlands and Northern Finland. *Suomen maataloustieteellisen seuran tiedote* , **21**, 1-7.

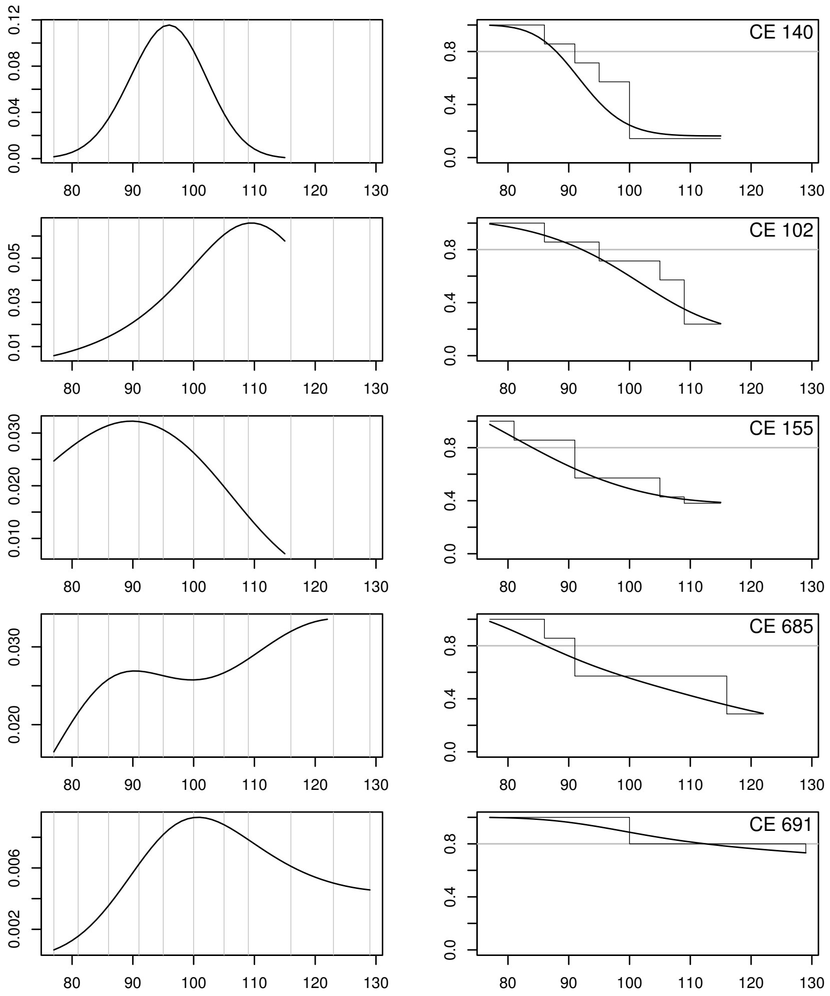

FIGURE 1. Left side: Estimated smooth hazard curves for five selected genotypes over days after planting (DAP). Vertical lines indicate the observation time points. Right side: Empirical (thin line) and fitted survival (thick line) for five selected genotypes over DAP. Grey line indicates 80 % survival.