AICOS Technologies AG
ADVANCED INDUSTRIAL CONSULTING IN OPERATIONS RESEARCH AND STATISTICS

Christian Keller, Dr. Yves-L. Grize
Efringerstrasse 32
CH-4057 Basel, Switzerland
Tel.: +41-61-686 98 77
Fax: +41-61-686 98 88
Web: http://www.aicos.com

Analysis of Process Data with Regression Trees

A project of AICOS Technologies AG

Introduction

For many production processes data measurements are more and more taking place automatically. This availability of data enables one to study relationships between process parameters in order to identify the causes of process deviations and to optimize process output.

The growing size and the complex structure of the data sets involved represent however serious challenges to the statistical methods. Missing values and outliers are often to be expected in production data. Moreover, complicated interactions between process parameters are likely to be present. Nevertheless relevant patterns in the data must be found and the dependence between process parameters appropriately modelled. Modern data mining techniques allow to solve such problems. In particular, the so-called classification and regression trees method (CART) turns out to be suited to production data.

What are classification and regression trees?

Classification and regression trees is a flexible method to summarize complex multivariate data sets and formulate simple prediction rules. The aim is to model the effect of a set of predictor variables on a response variable. If the response is measured on a continuous scale one speaks of a regression tree. For a categorical response on the other hand, the term classification tree is used.

Compared to classical statistical methods the advantages of classification and regression trees are the following:

- They can model linear as well as non-linear relations between predictor variables and the response.
- Interactions between particular variables are considered automatically and do not have to be specified in advance.
- Missing values are handled adequately.
- Only those variables are chosen for the final model which are important in predicting the response. An automatic data-driven variable selection is thus carried out.
- The results are easy to interpret even without profound statistical knowledge.

Classification and regression trees can be fitted with the statistical software S-PLUS using the function `tree`. All models can be specified with the flexible formula notation of S-PLUS. Numerous other functions serve for model selection and model checking purposes as well as for the visualization of the results (cf. figure 2).

A practical application: Analysis of a drying process

In a chemical product the variability of ethylester was too large with regard to the current specifications. The critical phase under investigation was the drying process. Drying serves as a reduction of the content of ethylester and of other substances in the product. The process is composed of three phases, the length of which depends on numerous factors, as for example the wet weight of the batch at the beginning of drying. Only an optimal choice of drying temperature, duration and other parameters enable to reduce the final content of ethylester below an acceptable value.

The purpose of the statistical analysis was to identify the causes for quality deviations and to build models that enable the formulation of recommendations on how variability can be reduced and process quality improved. Data for a total of 275 batches was available. Apart from the content of ethylester 28 process variables were measured.

In the first part of the analysis the data was explored with some selected visualization techniques (cf. figure 1). This exploratory data analysis helps to identify process disturbances and to interprete the results obtained from fitting statistical models.

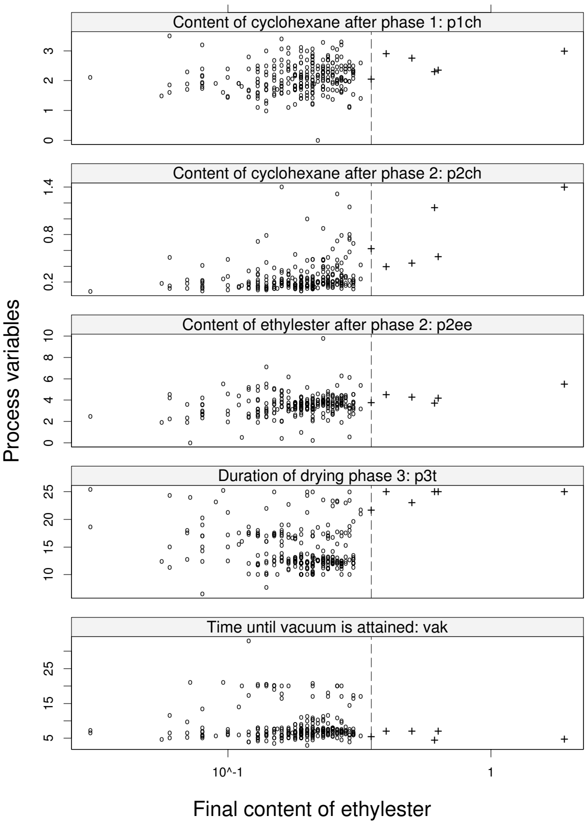

FIGURE 1: *Scatterplot for the final content of ethylester versus serveral process variables. The vertical reference line is the upper specification limit.*

The fitted regression tree for the final content of ethylester is displayed in figure 2. It can be interpreted as follows: At each node, represented as an oval, a regression rule for the response based on one predictor variable is formulated. For the tree in figure 2 the first rule is based on the content of cyclohexane after drying phase 2 (variable p2ch). For p2ch > 1.35% a high final content of 1% is predicted for ethylester. In the other case (p2ch < 1.35%) the mean content of ethylester after the drying process is only 0.2%. The numbers printed in the rectangular boxes, the leaves of the tree, are the predicted values for the response variable. For example, if the drying phase 3 (variable p3t) takes longer than 14.3 hours and the content of cyclohexane is below 0.17% after phase 2, then a low content of ethylester of 0.1% is predicted.

From the tree it follows that batches having a high content of cyclohexane after phase 2 tend to have a high final value of ethylester as well. If this relationship is of causal nature, the reduction of the content of cyclohexane must therefore already take place in phase 2.

Another look at figure 2 reveals that further important variables in the model are the content of cyclohexane after phase 1 (p1ch), the time until a certain vacuum is attained (vak), and the categorical variable f1d which indicates whether the exceeding of the standard temperature in phase 1 lasted for a short or a long time.

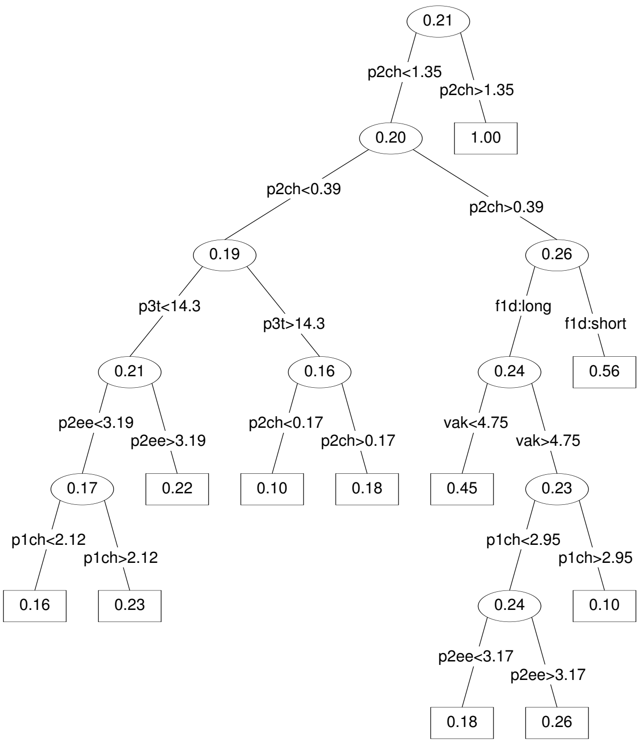

FIGURE 2: *Fitted regression tree for the final content of ethylester.*

Refinements of the methodology

Since tree size, i.e. the number of leaves, is not limited during the tree construction process, it may occur that a too complex model is fitted. Such models run the risk of leading to bad predictions of new observations, although they perfectly summarize the data used for model estimation. Therefore, the size of the tree must be selected carefully. For that purpose the established methodology is tree *cost-complexity pruning* and *cross-validation*. Both techniques can be implemented easily in S-PLUS with the functions `prune.tree` and `cv.tree`.

For cross-validation the data set is first divided into approximately ten sets of equal size. The first nine parts are called training data and the last part is the test data. Then a regression tree based on the training data is fitted in its maximal size. Afterwards that subtree, obtained by successively snipping of the least important splits (*pruning*), is selected which leads to the best prediction for the test data.

Summary

With the help of a regression tree the relationship between process parameters and the final content of ethylester could be described suitably. The results pointed out how to modify the drying process to reduce the variability and meet specifications.

Since strong interactions existed between some process variables, the use of regression trees proved to be particularly successful.