RIVM report 722601 007/2002

Correcting air pollution time series for meteorological variability
With an application to regional PM_{10} concentrations

H. Visser and H. Noordijk

This investigation has been performed by order and for the account of RIVM, within the framework of project S/722601, Measuring and Modelling.

RIVM, P.O. Box 1, 3720 BA Bilthoven, telephone: 31 - 30 - 274 91 11; telefax: 31 - 30 - 274 29 71

Abstract

It is well-known that a large part of the year-to–year variation in annual distribution of daily concentrations of air pollutants is due to fluctuations in the frequency and severity of meteorological conditions. This variability makes it difficult to estimate the effectiveness of emission control strategies.

In this report we have demonstrated how a series of binary decision rules, known as Classification And Regression Trees (CART), can be used to calculate pollution concentrations that are standardized to levels expected to occur under a fixed (reference) set of meteorological conditions. Such meteo-corrected concentration measures can then be used to identify 'underlying' air quality trends resulting from changes in emissions that may otherwise be difficult to distinguish due to the interfering effects of unusual weather patterns.

The examples here concern air pollution data (daily concentrations of SO_2 and PM_{10}). However, the methodology could very well be applied to water and soil applications. Classification trees, where the response variable is *categorical*, have important applications in the field of public health. Furthermore, Regression Trees, which have a *continuous* response variable, are very well suited for situations where physically oriented models explain (part of) the variability in the response variable. Here, CART analysis and physically oriented models are not exclusive but *complementary* tools.

Contents

Samenvatting

Dag-op-dag-variaties in meteorologische condities zijn een belangrijke oorzaak van variaties in het concentratieverloop van luchtveronreinigende stoffen. Deze aan meteorologie gekoppelde variaties werken ook door in *jaargemiddelde* concentraties. Daarom is het moeilijk om te beoordelen in hoeverre jaargemiddelde patronen van luchtverontreinigende componenten beïnvloed worden door emissiereducties. Zo'n beoordeling is zeer beleids-relevant omdat emissiereducties over het algemeen gepaard gaan met hoge kosten. Daarom zal er, om een een maatschapppelijke draagvlak te garanderen, een relatie gelegd moeten worden tussen trends in antropogene emissies enerzijds en trends in concentraties anderzijds.

In dit rapport tonen we aan hoe met behulp van een reeks binaire beslisregels, bekend staand onder de naam Classificatie- en Regressiebomen (Eng: CART), gemeten concentraties getransformeerd kunnen worden naar concentraties die er zouden zijn geweest onder standaard meteorologische condities. Deze meteo-gecorrigeerde concentraties kunnen vervolgens gebruikt worden om trends in luchtkwaliteit beter te identificeren.

CART-analyse en meer specifiek Regressieboom-analyse, heeft een aantal voordelen boven andere statistische technieken. In de eerste plaats is de methode parametervrij. Dat wil zeggen dat er geen aannames hoeven te worden gedaan over onderliggende kansverdelingen. In de tweede plaats mogen de relaties tussen een responsvariabele (concentratie van stof X) en de predictors (variabelen zoals temperatuur, neerslag, windrichting of windsnelheid) in hoge mate niet-lineair zijn. In de derde plaats zijn resultaten van een CART-analyse relatief eenvoudig te interpreteren.

Hoewel de voorbeelden in dit rapport gericht zijn op luchtverontreiniging, is de methode zeer geschikt voor andere milieuvelden zoals water- en bodemverontreiniging. Regressiebomen, waar de responsvariabele *continu* is, zijn van ook van belang voor situaties waar fysisch-georiënteerde modellen een deel van de variabiliteit van de responsvariabele beschrijven. Regressieboom-analyse is complementair aan deze fysische modellen. Classificatiebomen, waar de responsevariabele *nominaal* is, hebben grote relevantie voor het onderzoeksgebied van de Volksgezondheid.

We hebben verfijningen ontwikkeld voor de Regressieboom-benadering zoals beschreven door Dekkers en Noordijk (1997). Deze verfijningen omvatten: (i) controle van de data op uitbijters, ontbrekende waarnemingen en hoge correlaties tussen de predictors, (ii) controle op de responsvariabele of transformaties nodig zijn, (iii) cross-validatie van de geschatte Regressieboom, en (iv) evaluatie van de voorspelkracht van een Regressieboom in vergelijking met alternatieve voorspelmethoden.

Als een *case study* hebben we de methodologie toegepast op metingen van fijnstof (PM_{10}). We hebben 9 regionale stations geanalyseerd met metingen over de periode 1992-2001. De resultaten laten zien dat:

- regressieboom-modellen die gebaseerd zijn op *maandgemiddelde* concentraties, beter voldoen dan modellen op *daggemiddelde* waarden;
- langjarige trends in concentraties niet beïnvloed worden door variaties in meteorologie;
- concentraties een dalende tendens vertonen, net als emissies. Na correctie voor natuurlijke emissiebronnen (zeezout, opwaaiend stof en de achtergrondconcentratie van het Noordelijk halfrond) blijken emissies sterker te dalen dan concentraties.

Summary

It is well-known that part of the year-to–year variation in annual distribution of daily concentrations of air pollution in ambient air is due to fluctuations in the frequency and severity of meteorological conditions. This variability makes it difficult to estimate the effectiveness of emission control strategies.

In this report we have demonstrated how a series of binary decision rules, known as Classification And Regression Trees (CART), can be used to calculate pollution concentrations that are standardized to levels that would be expected to occur under a fixed (reference) set of meteorological conditions. Such adjusted concentration measures can then be used to identify 'underlying' air quality trends resulting from changes in emissions that may otherwise be difficult to distinguish due to the interfering effects of unusual weather patterns.

CART analysis, and more specifically Regression Tree analysis, has a number of advantages over other classification methods, such as Multiple Regression. First, it is inherently non-parametric. In other words, no assumptions have to be made a priori regarding the underlying distribution of values of the response variable or predictor variables. Second, the relationship between the response variable (concentration of a pollutant) and predictors (meteorological variables) may be highly non-linear. Third, the estimation results of a CART analysis are relatively easy to interpret.

Although the examples given here concern air pollution, the methodology could very well be used for water and soil applications. Regression Trees, having a *continuous* response variable, are very well suited to situations where physically oriented models explain (part of) the variability of the response variable. Here, Regression Trees and physically oriented models are not exclusive but *complementary* tools. Furthermore, Classification Trees, where the response variable is *categorical*, have important applications in the field of Public Health.

We have refined the methodology of Dekkers and Noordijk (1997) for Regression Trees. Refinements comprise (i) checks on the data for outliers, missing values and multi-collinearity among the predictors, (ii) checks for transformation of concentrations prior to the estimation of a Regression Tree, (iii) cross-validation of the final optimal tree and (iv) evaluation of the predictive power of the final tree in relation to alternative (rival) models.

The Regression Tree methodology, applied as a case study to nine regional stations of PM_{10} in the Netherlands, has yielded the following results:

- RT models based on *monthly* concentrations outperformed those based on *daily* data. Apparently, monthly averaged meteorology is more influenced by large-scale meteorology in Europe, governing periods with extreme concentrations;
- Long-term trends in PM_{10} concentrations have not been not influenced by meteorological variability.
- Regional concentration trends show large similarities to trends in *emissions*. If we correct concentrations for natural emission sources (sea salt, wind-blown dust and northern hemisphere background concentrations), emissions decrease faster than concentrations.

1. Introduction

1.1 Why statistical modelling?

A wide diversity of mathematical models is applied within the RIVM Office for Environmental Assessment (MNP in Dutch). These models are based on physical, chemical, meteorological and biological relationships. Reality is approximated as well as possible in these models. What 'as well as possible' means, can be verified by environmental measurements.

We will denote the model-based approach here as 'white box modelling'. In many cases *white box modelling* will give a deterministic approach to reality. As an example we can cite the OPS model (Jaarsveld, 1995) by which the dispersion and deposition of a number of air pollution components are modelled as a function of emissions and meteorology.

Contrary to *white box modelling* we also have so-called 'black box modelling'. With this second approach we mean the modelling of measurements on the basis of statistical principles. A measurement series is seen as a deterministic signal and a stochastic residual signal, the 'noise'. A series of measurements can therefore be viewed as a *possible realization* of reality. Some slightly different outcomes would have been equally likely. We will denote these series as 'time series'.

Within the statistical approach, relationships (or associations) are estimated by calculating correlation. Correlation points to the similarity in patterns, but does not prove causality. Therefore, we use the term 'black box'. Examples of *black box modelling* are illustrated in Multiple Regression models, ARIMA models or methods based on the Regression Tree, described in this report.

The mixture of *white box* and *black box modelling* is called *grey box modelling.* If a physical model only partly describes reality, we can describe the 'difference' between *white box model predictions* and reality (the measurements) by statistical modelling.

An example of *grey box modelling* is the modelling of PM_{10}-concentrations in the Netherlands by the OPS-model. The OPS-model describes the anthropogenic contribution to measured PM_{10} concentrations. However, by doing so, only half the concentration variations are explained. Research has shown that the 'rest' is largely explained by the share of natural sources (Visser et al., 2001). From this moment, we could describe and predict the difference between measurements and OPS-predictions statistically, using such geostatistical techniques as Universal Kriging.

In the absence of a physical model, it is clear that statistical models can play an important role in finding relevant relations. But even with a physical description of our environmental reality, statistical modelling can be of importance. Statistical inferences are pre-eminently suited as a diagnostic tool for unraveling and explaining differences between measurements and the predictions of physically oriented models. Three examples follow:

1) The correction of environmental measurements for meteorological conditions. Dekkers and Noordijk (1997) describe a generic method based on Regression Trees by which the time series of air pollution components (PM_{10}, SO_2, NO_x, O_3, and black smoke) is corrected for meteorological conditions. Such a correction is of great importance for policy-makers because after correction one can make inferences on the influence of emission reductions on the actual ambient concentrations with much more certainty.

Statistical techniques such as Regression Tree analysis or Multiple Regression yield insights, which may be defined later in terms of physical relationships. An example is given in Chapter 5 for the time series of PM_{10} concentrations. It appears that months with extreme droughts and cold periods correspond to months with very high PM_{10} concentrations. This sort of relationship has not been taken into account in RIVM models such as OPS or EUROS, but could be formulated into a physical framework and added to these models.

2) A second example is found in the field of forecasting. For components such as PM_{10} and O_3 RIVM produces smog forecasts on television (Teletekst) 1 and 2 days in advance. Simple statistical models, such as the autoregressive model with two parameters, have often been found to outperform forecasts of complex white models. Here, forecasts of black and white models could be evaluated on a test set, where the best method is declared the winner. Or one might choose to *combine* forecasts of both models (Makridakis et al., 1982).

3) A third example of great importance to the MNP, is the sensitivity analysis of complex computational-intensive models (Saltelli et al., 2000). One method in this field is to find the sensitivity of a certain output variable, y, to variations in the input variables, $\mathbf{x} = (x_1, x_2, \ldots, \text{to } x_m)$. Which x_i has the largest impact and which the lowest? If the computational time of one model run is in the order of hours or even days, this question is not trivial. Now, if we calculate y for a limited number of input combinations **x**, we can estimate a Regression Tree model or a Multiple Regression model between these sets (y,**x**). Once having estimated a black box model, one can easily deduce an ordering in importance of the various input variables.

Such an exercise has been performed for the RIVM model PcDitch, consisting of 100 response variables and a set of 200 predictors. Screening the 200 predictors by Regression Tree analysis yielded a set of 20 predictors governing specific response variables.

Physically oriented models and statistical models such as Regression Trees, are not rivals but complementary tools in explaining what we know and what we don't know about our data (Photo: H. Visser).

1.2 The need for meteo corrections

The analysis of a series of measurements is generally done to show trends in concentrations that are influenced by human activities. The point is to clarify how society influences the ambient air pollution concentrations. On the one hand, the ongoing economic growth leads to more production and emissions, while, on the other, environmental abatement policies and the abatement of emissions leads to mitigation of concentrations. A long-term time series of measurements may give us insight into the influence of the economy and environmental policy.

A third influence on the ambient concentrations is that of the meteorology. The influence of the weather is not constant in time. For example, taking particulate matter (PM) and SO_2, we know that a cold winter in the Netherlands will yield higher average concentrations. These higher PM concentrations are partly the result of more transport from abroad, as sub-zero temperatures are usually accompanied by a continental air flow that is generally more polluted than air masses from the Atlantic Ocean. During a period of frost the general dispersion conditions are such that inversions occur more often. Furthermore, emissions tend to be higher during these periods as heating of homes consumes more energy and cold starts of cars produce more pollution. During a long cold spell high wind speeds may also re-suspend airborne crustal material from the barren and dry fields. A year with more rain than usual will lead to lower concentrations of PM, mainly because of either rainout or washout, and to a lesser extent because it is harder for crustal material to become re-suspended when soils are moist.

In this report we will demonstrate how a series of binary decision rules, known as Classification And Regresion Trees (CART), can be used to calculate pollution concentrations that are standardized to levels that would be expected to occur under a fixed (reference) set of meteorological conditions. Such meteo-corrected concentration measures can then be used to identify ‘underlying’ air quality trends resulting from changes in emissions that may otherwise be difficult to distinguish due to the interfering effects of unusual weather patterns.

1.3 Regression Tree analysis

A method has been developed at the RIVM to analyse the influence of meteorology on the trends in air pollution (Dekkers and Noordijk, 1997). This method uses the ambient air pollution concentrations in the Netherlands in combination with daily-averaged meteorological values. The meteorological factors influencing concentrations are divided into different classes by way of *Regression Trees* (Breiman et al., 1984). Regression Trees comprise part of the so-called Classification and Regression Trees (CART) methodology. The difference between Classification Trees and Regression Trees is that the response variable is categorical in Classification Trees and continuous in Regression Trees.

Regression Tree analysis (RTA) is a statistical technique that divides the set of measurements into two sub-sets on the basis of meteorological criteria. The criterion for the division of the sub-sets is the minimization of the variance of the two sub-sets. After this first step in the analysis, one of the sub-sets is itself again divided into two new sub-sets, with again the

criterion of minimization of variance. Eventually, this leads to a 'tree' of classes describing the influence of meteorology on the concentrations.

Once we have generated a tree, we want to check if all the nodes in the tree are needed or if we should *prune* the tree. The rationale for pruning the tree is that we want to have a model for our data that is as *parsimonious* as possible, while keeping certain desirable characteristics in tact (such as the predictive power of the tree, see below).

The final nodes of the tree are called 'leaves'. By averaging all concentrations on days that correspond to that specific leaf we get an RT prediction for days that fall in the particular meteo class. It should be noted that the term 'predictions' is used in the common statistical sense, i.e. we may predict both values *within* the time series available and predict the future (forecasting).

The 'predictive power' of the tree is found by calculating the mean squared error of the prediction errors. The prediction error stands for the difference between the actual concentration on a specific day and the corresponding leaf prediction (the average value of all concentrations of falling in that leaf).

Once a suitable tree has been estimated, we want to correct annual averaged concentrations or annual percentiles for meteorological conditions. Basically, there are two approaches. The first approach has been proposed by Stoeckenius (1991). For every year the frequency of occurrence of a meteo class is determined, the frequency of occurrence determines the actual value of the correction factor. As an example, when the meteo class of tropical days normally occurs three times a year and in a specific year there are six of these days, the calculation of these tropical days in the yearly average is less than average and the correction factor becomes 0.5. However, when by chance a certain meteo class does not occur, an estimate is made of the expected concentrations by using concentrations from other years in that specific meteo class. We will describe an expanded version of this procedure in more detail in Noordijk and Visser (in prep.).

A second approach is to estimate the mean concentration μ_y for all concentrations y_t . If we denote the predicted concentration at time t as $\hat{y}_t$, then we define the particular concentrations due to meteorological conditions as the difference between $\hat{y}_t$ and μ_y. Finally, the meteo-corrected concentration on day t is $y_{corr,t} = y_t - (\hat{y}_t - \mu_y)$. Now, annual averaged concentrations or percentiles are simply calculated on the corrected daily data $y_{corr,t}$. In this document we will apply this correction approach, while in Noordijk and Visser (in prep.) we will evaluate both meteo-correction methods.

CART analysis, and more specifically Regression Tree analysis, has been applied in many fields. For references see Ripley (1996) and Venables and Ripley (1997). Many references can by found via search path *CART 'Regression Tree*'*. However, Regression Tree analysis has not been applied as frequently in air pollution research. Two references on application to ozone data are Stoeckenius (1991) and Gardner and Dorling (2000). In the latter article Regression Tree analysis, Linear Regression and Multilayer Perceptron Neural Networks are compared using hourly surface ozone concentrations.

1.4 Refinements

We apply Regression Tree analysis to time series of air pollutants. By doing so, we have to ensure that our final tree model:

- fits the data well;
- is physically plausible;
- is able to withstand a comparison to alternative (rival) models.

Harvey (1989, pages 13-14) summarizes the criteria for a good model as proposed in the econometric literature. These criteria equally hold for environmental time-series models and, more specifically, for Regression Tree models.

The following six criteria from Harvey (1989) represent an elaboration of the three points above:

(a) Parsimony. A parsimonious model is one that contains a relatively small number of parameters. Other things being equal, a simpler model is preferred to a complicated one. In general, there can be considerable advantages in starting with a general model and then simplifying it on the basis of statistical tests.

(b) Data coherence. Diagnostic checks are performed to see if the model is consistent with the data. The essential point is that the model should provide a good fit to the data and the residual, be relatively small, and approximately random.

(c) Consistency with prior knowledge. The size and magnitude of parameters in the model should be consistent with prior information. And the same should hold for the classification rules from a specific Regression Tree. This information should relate to a physical, chemical or meteorological context.

(d) Data admissibility. A model should be unable to predict values which violate definitional constraints. For example, concentrations cannot be negative.

(e) Validation (structural stability). The model should provide a good fit, inside and outside the sample. In order for this to be possible, the parameters or classification rules should be constant within the sample period and this constancy should carry over to data not in the sample period. The latter data could fall within the timespan of the sample (principle of cross validation) or could lie in the future, the post-sample period.

(f) Encompassing. A model is said to encompass a rival formulation if it can explain the results given by the rival formulation. If this is the case, the rival model contains no information that might be of use in improving the preferred model. In order fulfil its encompassing role, a model does not need to be more general than its rivals.

To meet these criteria as best as possible, we have made a number of refinements to the approach given by Dekkers and Noordijk (1997). These refinements comprise diagnostic checks to the data set prior to the estimation of a specific Regression Tree, transformation of data, analysis of residuals, testing the stability of the final tree by the principle of cross-validation, and comparison of Regression-Tree performance to alternative/rival models. We will describe these refinements in detail in §2.4.

1.5 The report

In Chapter 2 we will describe the Regression Tree approach in more detail, illustrating the theoretical considerations with an example on SO_2 concentrations. Chapter 3 will be devoted to a short, in-depth, description on the implementation of a Regression Tree analysis in S-PLUS. S-PLUS is the standard statistical software package of RIVM (Dekkers, 2001). More details on this topic will be given in Visser (in prep.). Our approach has resulted in an eight-step procedure, summarized in §2.9. In Chapter 4 we will illustrate this eight-step procedure by use of a simulation example. Simulated examples have the advantage that the correct solutions are known a-priori and that the estimation procedure can be judged on its merits.

In Chapters 5 and 6 we will give case studies using the methodology given in Chapter 2. Chapter 5 deals with nine regional PM_{10} stations in the Netherlands. In Chapter 6 we will analyse 22 regional SO_2 stations. For both pollutants we will show that meteo-corrected regional concentrations show the same long-term trend as the uncorrected data. The extreme concentrations in 1996, and to a lesser extent 1997, appear to originate from unusual cold and dry winters at that time.

2. Regression Tree analysis

2.1 Method

Tree-based modelling is an exploratory technique for uncovering structure in data and is increasingly used for:

- devising prediction rules that can be rapidly and repeatedly evaluated;
- screening variables;
- assessing the adequacy of linear models;
- summarizing large multivariate data sets.

Tree-based models are useful for solving both classification and regression problems. In these problems, there is a set of classification or predictor variables, $\mathbf{x} = (x_1, x_2, \ldots, x_m)$, and a single-response variable, y. In the literature tree-based models are denoted by the abbreviation CART (Classification And Regression Trees).

If y has *discrete values*, classification rules take the form:

> If $x_1 < 2.3$ and $x_3 \in \{A,B\}$
> then y is most likely to be in level 5

If y is *numeric*, regression rules for description or prediction take the form:

> If $x_2 < 2.3$ and $x_9 \in \{C,D,F\}$ and $x_5 \geq 3.5$
> then the predicted value of y is 4.75

In the first case we speak of a *classification tree*, and in the second, of a *regression tree*. A classification or regression tree is the collection of many such rules displayed in the form of a binary tree, hence the name. The rules are determined by a procedure known as *recursive partitioning*.

Tree-based models provide an alternative to linear and additive models for regression problems, and to linear and additive logistic models for classification problems.

Compared to linear and additive models, tree-based models have the following advantages:

- easier interpretation of results when the predictors are a mix of numeric variables and discrete variables;
- invariant to monotone re-expressions of predictor variables;
- more satisfactory treatment of missing values;
- more adept at capturing non-additive behaviour;
- allow more general interactions (i.e. as opposed to particular multiplicative form) between predictor variables;
- can model response variables, y, having more than two levels.

In the following example (in italic) we deal only with regression trees, applied to air-pollution time series. In the second paragraph (in italic) we show a *hypothetical* air-pollution example. The data apply to a station measuring pollutant X. Concentrations are expressed in $\mu g/m^3$.

> ***A factory emitting a substance X, is situated in the vicinity of the station (direction north-northeast or more precise: 30 degrees). If the daily averaged wind direction is between 5 and 55 degrees, we measure a pollutant concentration of 50 $\mu g/m^3$ (the spread in wind direction is due to dispersion of the plume; concentrations are assumed to be homogeneous over the full width of the plume). For all other directions, we measure a constant background concentration of 20 $\mu g/m^3$.***
>
> ***To generate the hypothetical concentration series X, we use real daily wind data for the Netherlands over the period 1992-2001. Thus, we have 3653 daily concentration of either 20 or 50 $\mu g/m^3$, depending only on the specific wind direction on that day. For estimating the Regression Tree, we add 11 other meteorological variables as well as 'wind direction', making a total set of 12 predictors (see data at the beginning of Chapter 4).***

Clearly, the relationship between concentrations X and meteorology is simple and highly non-linear. Below we estimate a Regression Tree for X and the 12 predictors. The result is shown in **Figure 1.**

Figures within ellipses (nodes) or rectangular boxes (final nodes or *leaves*) represent averages of the concentrations falling under meteo conditions described above the nodes. Figures beneath the nodes are deviances, i.e. the sum of squared deviations of concentrations and the average of all concentrations falling in that node. If the sum of deviances in the leaves is much lower than the deviance in the highest node in the tree, we have made great improvement by adding meteorological factors.

From the Regression Tree in Figure 1 we see that the tree algorithm identified the right variable (Windr) to make splits for: wind direction. Furthermore, the values found for the wind sector [5.0,55.0] degrees are reasonably accurate: [7.5,55.5] degrees. Finally, if wind direction is found in the sector [7.5,55.5], concentrations are perfectly predicted: 50 $\mu g/m^3$ (deviance equals zero). The same holds for wind directions > 55.5: 20 $\mu g/m^3$. If wind directions are < 7.5 degrees, predictions are slight to high: 28.4 $\mu g/m^3$.

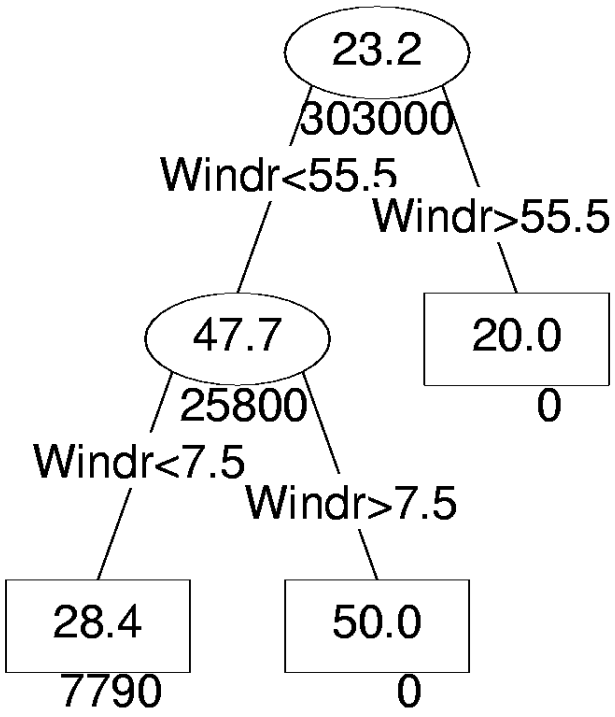

Figure 1 *Regression Tree for a hypothetical pollutant X, with increased concentrations for wind directions from 5 to 55 degrees.*
Figures in the ellipses (nodes) represent the average of all concentrations falling under a specific meteo condition and mentioned above the node. Rectangular boxes or final nodes are called the *leaves* of the tree. The figures beneath nodes and leaves are deviances, i.e. the sum of squared deviations of concentrations in that node/leaf and the average of concentrations in that node/leaf. The splitting process depends on the reduction of the initial deviance (here 303,000 $[\mu g/m^3]^2$) to a much smaller deviance. The sum of deviances in the leaves accounts for: 7790 + 0 + 0= 7790 $[\mu g/m^3]^2$.

For the exact definition of terms such as 'regression tree predictions', 'deviance' and 'predictive power', we need some formal definitions. These are taken from the basic book on regression trees by Breiman et al. (1984). We will use these definitions in §2.3 and §2.7 (yellow text blocks).

Definitions I

As mentioned above, for estimating a Regression Tree we have a response variable y_i (concentration of some pollutant) and m predictor variables $\mathbf{x_i} = (x_{1,i}, x_{2,i}, \ldots, x_{m,i})$. In this report $\mathbf{x_i}$ will consist of a set of m meteorological variables. The pointer i may be expressed in days, as in Figure 1, or in months, as in Chapter 5.

Suppose all days or months fall into J classes (the leaves of the tree). We name these classes 1,2, …, J. The estimated Regression Tree (RT) defines a *classifier* or *classification rule* $c(\mathbf{x_i})$, which is defined on all $\mathbf{x_i}$, i =1, 2, … , N. The function $c(\mathbf{x_i})$ is equal to one of the numbers 1,2, … , J.

Another way of looking at the classifier c is to define A_j , the subset of all $\mathbf{x_i}$ on which $c(\mathbf{x_i}) = j$:

$$A_j = \{ \mathbf{x_i} \mid c(\mathbf{x_i}) = j \} \qquad (1)$$

The sets A_j, j = 1, 2, … , J are all disjunct and the unity of all sets A_j spans exactly all the N cases we have.

2.2 Pruning

Trees with too many nodes will over-fit our data. In fact, we could construct a tree with as many nodes as days or months we have. The fit to the data will be very good. However, such a tree does not serve our goal of finding parsimonious models (criterion (a) in section 1.4). Therefore we have to look for an analogue of variable selection in Multiple Regression analysis.

The established methodology is tree cost-complexity pruning, first introduced by Breiman et al. (1984). They considered rooted subtrees of the tree T grown using the construction algorithm, i.e. the possible result of snipping off terminal subtrees on T. The pruning process chooses one of the rooted subtrees. Let R_i be a measure evaluated at the leaves, such as the deviance (compare equation (3)), and let R be the value for the tree, the sum over the leaves of R_i. Let the size of the tree be the number of leaves.

By pruning Regression Trees, we try to find an optimal balance between the fit to the data and the principle of parsimony (criterion (a) in Section 1.4). To this end the full tree is pruned to a great number of subtrees and evaluated as for minimization of the so-called cost-complexity measure. *Photo: H. Visser*

Then, Breiman et al. showed that the set of rooted subtrees of T which minimize the cost-complexity measure:

$R_\alpha = R + \alpha \, . \, size$

is itself nested. In other words, as we increase α, we can obtain optimal trees through a sequence of snip operations on the current tree (just like pruning a real tree). For details and proofs see Ripley (1996).

The best way of pruning is using an independent test set for validation. We can now predict on that set and compute the deviance versus α for the pruned trees. Since this α will often have a minimum, and we can choose the smallest tree of which the deviance is close to the minimum. If no validation set is available, we can make one by splitting the training set. Suppose we split the training set into 10 (roughly) equally sized parts. We can then use 9 to grow the tree and the 10th to test it. This can be done in 10 ways and we can average the results. This procedure is followed in the software we have implemented in S-PLUS. Note that as ten trees must be grown, the process can be slow, and that the averaging is done for fixed α and not for fixed tree size.

We give an example for SO_2 concentrations at Posterholt station in the Netherlands. Concentrations consist of daily averages over the 1989 – 2001 period. The Regression Tree is shown in **Figure 2**. An example of a cross-validation plot for pruning is given in **Figure 3A**. The optimal tree size has been computed (values lower x-axis) for a series of α values (on upper x-axis). The values are averages of the ten-fold cross-validation procedure described above. The values on the y-axis are the deviances over all leaves of the selected trees. The graph has a clear minimum around six leaves. Thus, we decide to prune our full tree back to one with six leaves. The result is shown in **Figure 3B**.

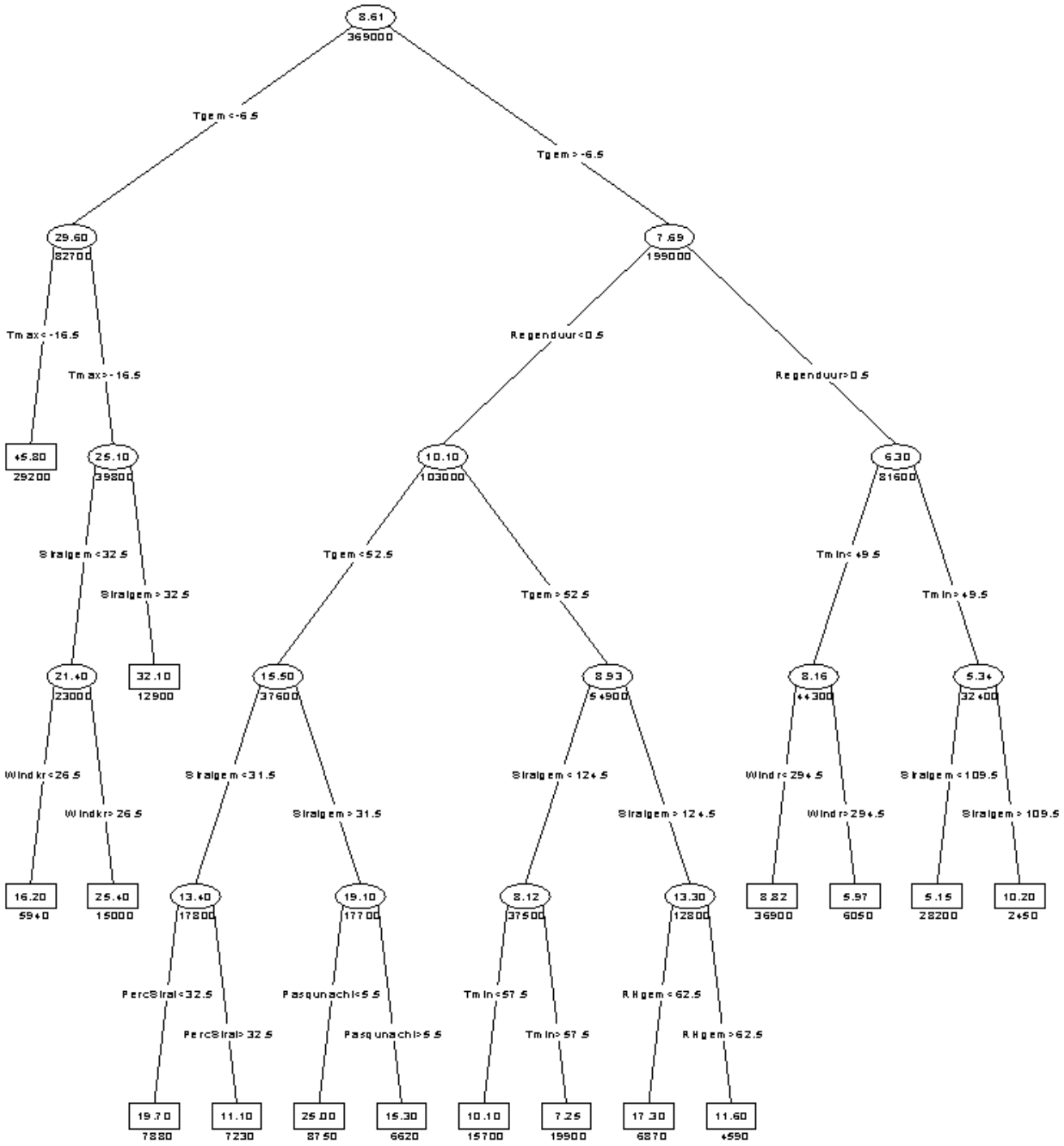

Figure 2 Regression tree for daily SO_2 concentrations measured at station Posterholt in the Netherlands. Sampling period covers 1989 through 2001.

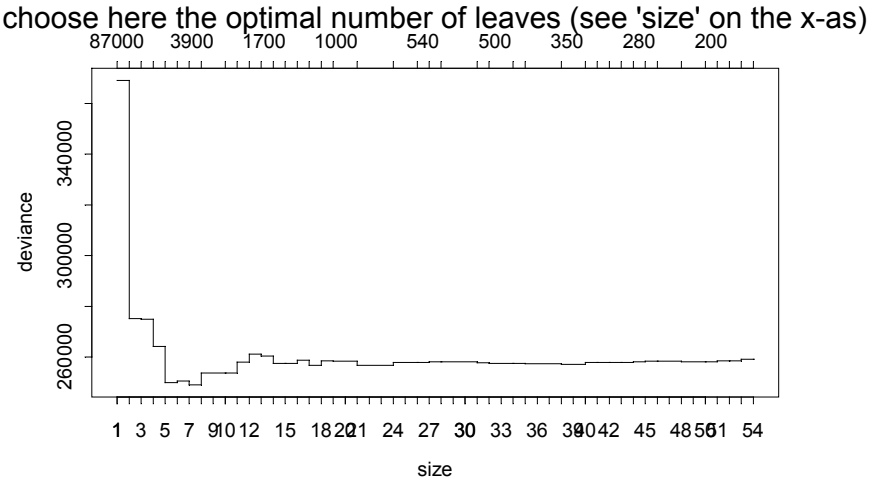

Figure 3A *Cross-validation plot for pruning the tree from Figure 2. Minimum is around tree size 6.*

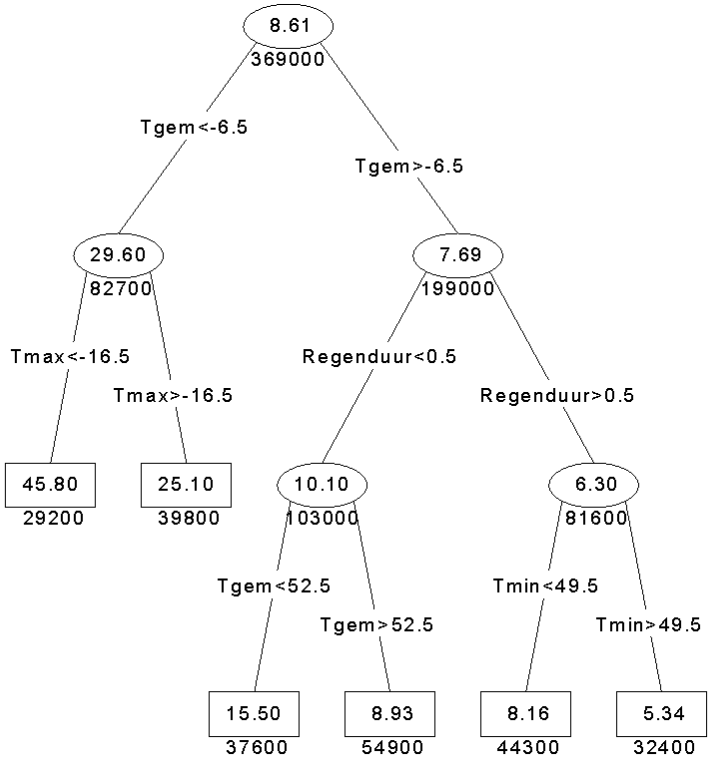

Figure 3B *Pruned tree for daily SO_2 concentrations at Posterholt. Temperature variables Tgem, Tmax and Tmin are expressed in 0.1 ˚C. Variable Regenduur stands for rain duration (in hours).*

2.3 Predictions

Given an optimal pruned tree we can calculate daily or monthly predictions for concentrations by averaging all concentrations that fall in a certain meteo class (leaf). Therefore if we have meteo conditions which fall in that meteo class, our prediction is simply the leaf average. Thus we have as many prediction values as leaves. The formal definitions are given below.

Definitions II

Given an estimated tree with m leaves (or meteo classes), we want to make a single prediction for days or months, with meteo conditions applying to a specific leaf j. On the basis of the definitions in Section 2.1., the classification rule c puts each daily concentration in one of the meteo classes: 1,2, … , J. We can now define an RT prediction for all concentrations y_i falling in meteo class j. Normally, the arithmetic mean is taken for all concentrations falling in class j. Other values could be the median or the modus. In this report we will take the arithmetic mean.

Thus, the prediction $\hat{y}_{i,j} \equiv \mu_j$ for a concentration on day or month i, belonging to class j, is:

$$\mu_j = \frac{1}{N_j} \sum_{\{i|c(x_i)=j\}} y_i \qquad (2)$$

where N_j is the number of cases in class j.

The deviance, D_j, is a measure of the variation of concentrations falling in class j. It is defined as:

$$D_j = \sum_{\{i|c(x_i)=j\}} (y_i - \mu_j)^2 \qquad (3)$$

The predictions μ_j are given in Figure 1 within the ellipses and boxes, while the corresponding deviance is given below each ellipse or box.

As a general notation we will denote the predictor function by ‘d’. Thus, for each concentration y_i we obtain the prediction $\hat{y}_i = d(\mathbf{x_i})$, with $\hat{y}_i$ one of the numbers $\mu_1, \mu_2, \ldots, \mu_J$.

The variance of a specific prediction μ_j simply follows from the deviance (3): $var(\mu_j) = Dj/Nj$. By using this relationship we can generate confidence limits for predictions.

We will now follow the SO_2 example from the preceding sections. The daily concentrations y_i are plotted in **Figure 4** against the predictions $\hat{y}_i$. Because we have a classifier with six meteo classes, we have six values on the y-axis. These six values are identical to those given in the leaf boxes in Figure 3.

In interpreting the pruned in Figure 3B, we expect very high concentrations (45.8 $\mu g/m^3$) if daily averaged temperatures are below –0.65 ˚C **and** maximum daily temperatures are below 1.65 ˚C. This meteo class is typically a measure for winter smog conditions. We expect the lowest SO_2 concentrations (5.34 $\mu g/m^3$) if daily temperatures are above –0.65 °C **and** rain duration is over 0.5 hour **and** the minimum daily temperature is above 4.95 ˚C.

Figure 5 shows the concentrations and corresponding predictions as a time-series plot.

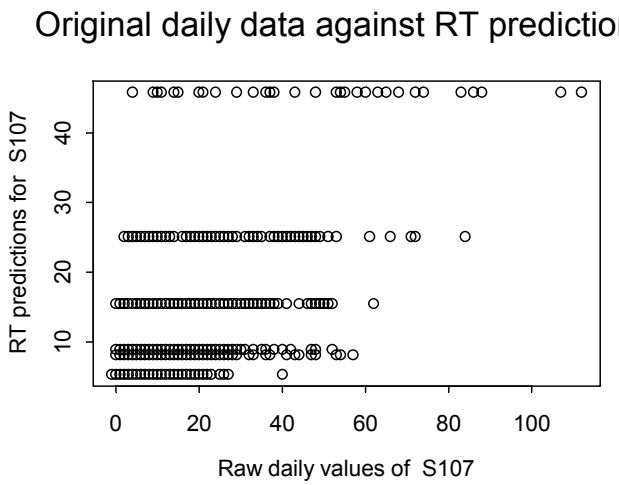

Figure 4 *Scatterplot of daily SO_2 concentrations for Posterholt station (x-axis) measured against RT predictions.*

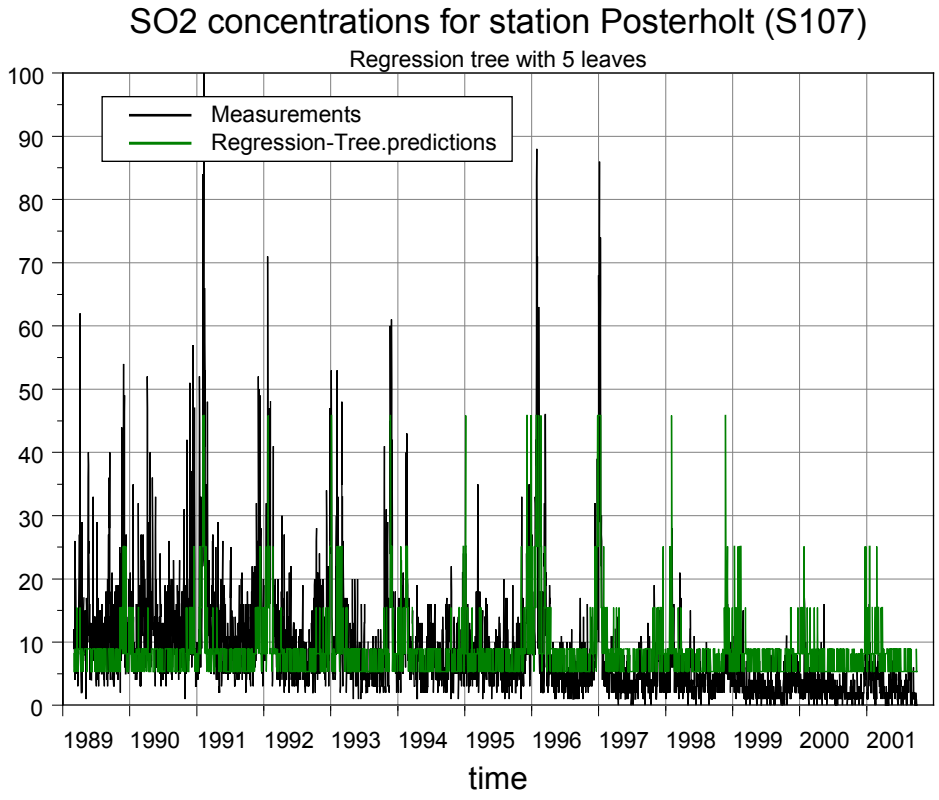

Figure 5 *Daily SO_2 concentrations for Posterholt station with RT predictions.*

2.4 Predictive power in relation to rival models

To compare two alternative models for the same concentration data, we need a measure for prediction accuracy. Here, a rival model could be a regression tree with more or less nodes than the current estimated tree. A rival model could also be a model following a different statistical approach. Examples are the overall-mean predictor, the naive or persistence predictor and the Multiple Regression predictor. The overall-mean predictor uses the overall mean of all concentrations as a constant prediction for all new cases (days or months). The naive predictor uses the value of the preceding day or month as a predictor for the following day or month.

In the following text box we will define simple indices to evaluate the predictive power of a specific estimated tree, relative to the three rival models mentioned above. All indices are based on quadratic prediction errors.

Definitions III

Breiman et al. (1984) defines the following *mean squared error* R_{RT}*(d) for a Regression Tree (RT) with prediction rule d:

$$R^*_{RT}(d) = \frac{1}{N}\sum_{i=1}^{N}[y_i - d(x_i)]^2 \qquad (4)$$

The asterisk is used to denote the fact that R_{RT} is *estimated* on the data. Another well-known measure for prediction performance is the minimum absolute deviation criterion (MAD). This criterion is less sensitive to outliers. However, throughout this report we will use only the criterion defined in (4).

In a similar manner we may define mean squared errors for rival models. In our approach we routinely estimate three rival models: the overall mean predictor, the naive predictor and the Multiple Regression predictor. The overall mean predictor simply uses the overall mean of the N concentrations (μ) as prediction for all days or months y_i. We denote its mean squared error by R_μ*. The naive predictor is also known as the persistence predictor and for each concentration y_i simply uses the concentration of the preceding day, thus y_{i-1}. We will denote its mean squared error by R_{naive}*. The third model is the well-known Multiple Regression model estimated on the same data set using the same meteorological predictor set $\mathbf{x_i}$. Its mean prediction error is denoted by R_{MR}*.

From the mean squared errors given above we define the performance of our RT model relative to that of these three rival model as:

$$RE^*_{\mu}(d) = \frac{R^*_{RT}(d)}{R^*_{\mu}} \qquad (5a)$$

$$RE^*_{naive}(d) = \frac{R^*_{RT}(d)}{R^*_{naive}} \qquad (5b)$$

$$RE^*_{MR}(d) = \frac{R^*_{RT}(d)}{R^*_{MR}} \qquad (5c)$$

We also express these relative prediction measures in percentages to express the improvement by using the RT model against one of the rival models:

$$P^*_{\mu}(d) = [1 - RE^*_{\mu}(d)] * 100 \quad (\%) \qquad (6a)$$

$$P^*_{naive}(d) = [1 - RE^*_{naive}(d)] * 100 \quad (\%) \qquad (6b)$$

$$P^*_{MR}(d) = [1 - RE^*_{MR}(d)] * 100 \quad (\%) \qquad (6c)$$

We note that in linear regression applications, the term $1 - RE^*_{\mu}(d)$ is called the variance explained by d. And the sample correlation, expressed as percentage, would be equal to $P^*_{\mu}(d)$. However, in general, $R_{RT}^*(d)$ in equation (4) is not a variance; it does not make sense to refer to it as 'the proportion of variance explained'. The same holds for the squared correlation.

As an example we give the indices defined above for the SO_2 regression tree shown in **Figure 3**. We find for this tree the following values: $R_{RT}^*(d) = 53$, $R_\mu^* = 81$ and $R_{naive}^* = 41$ $[\mu g/m^3]^2$. Now, the relative indices are $RE_\mu^*(d) = 0.65$, $RE_{naive}^*(d) = 1.28$, $P_\mu^*(d) = 35\%$ and $P_{naive}^*(d) = -28\%$.

Clearly, the RT predictions are better than the overall-mean predictor (35%). However, the RT predictions are 28% worse than those by the naive or persistence predictor. This indicates that we have to look for better trees to outperform this rival model (compare results in §2.8).

For the model shown in Figures 3B and 5, Regression Tree predictions are 28% worse than those made by the naive or persistence predictor (today's prediction is yesterday's value).
Photo: H. Visser

2.5 Transformation of concentrations prior to analysis

In time-series analysis it is good practice to check the data for trend and time-dependent variability (heteroscedasticity). By the latter we mean that the variability of concentrations may depend on the actual average level. Heteroscedasticity is tested through Range-Mean plots (compare to Figure 10). The most common way of removing heteroscedasticity is by applying a log-transformation prior to the estimation of a regression tree. Regression Tree predictions can be transformed back to the original scale by taking exponentials.

In Regression Tree analysis a trend is not part of the set of predictors **x**. If we were to add a linear trend to the set **x**, the Regression Tree procedure would split our time-series into a number of consecutive blocks over time (highest nodes in the tree). Then, for each block the dependence on meteorology will be accounted for (lower nodes in the tree). In this way, the number of days or months on which to estimate the actual meteo-related tree is reduced considerably. For this reason, we did not consider the addition of a trend to **x** [1)].

However, for estimating an RT, a particular time series should have *stationary* characteristics. Stationarity in time-series analysis means that:

- the true (local) mean of a series is time-independent. E.g., a series may not contain a long-term trend;
- the correlation structure is time-invariant. E.g., a stationary time series has a variance which is invariant (homoscedastic) over time.

As an alternative, we remove a clear trend in the data, if necessary, and analyse the trend-corrected data by Regression Tree analysis. Regression Tree predictions and meteo-corrected concentrations (compare §2.7) are transformed back to the original scale by using the inverse transformation.

Of course, there is a 'grey area', where it is not clear if we should remove the trend or not**. In these cases it is advisable to estimate a Regression Tree to both untransformed and transformed concentrations, and to check for differences between the two.**

We have implemented three transformations in the ART software (Chapter 3):

$$y_t' = \log (y_t + b) \qquad (7)$$

with 'b' a constant, such that $(y_t + b)$ is positive for all time steps t. This transformation stabilizes variability over time.

---

1) We note that the approach of adding a linear trend to the set of predictors also has an advantage. Because Regression Trees are estimated for two or more sub-periods, we can test the stability of the tree as a function of time.

If significant trends over time exist, we apply one of the following transformations:

$$y_t' = y_t - \text{trend}_t \quad (8a)$$

or

$$y_t' = y_t / \text{trend}_t \quad (8b)$$

Transformation (8b) has a stabilizing effect on the variability of the concentrations.

The trends in equations (8a) and (8b) are satisfactorily estimated by an n-degree polynomial $y_t = a_0 + a_1 t + \ldots + a_n t^n$. As for concentration data, n has a maximal value of 3. To ensure stable estimates of the polynomial, we use the function POLY from S-PLUS.

Another important group of transformations is formed by the Box-Cox transformations, which have advantages in transforming data to normality. However, a disadvantage is the interpretation of the transformed concentrations. We decided *not* to implement this group of transformations.

As an example we have transformed the Posterholt SO_2 data by using transformation (8b). This transformation removes the trend, and the trend dependence of the variance, at the same time. Now the tree shown in Figure 3 for untransformed data changes into the tree shown in **Figure 6**. Comparing **Figures 3** and **6** we see that the main variables used for splitting are similar. However, the values for splitting differ between trees. The splitting order is also different. We conclude that trend removal to be an essential first step in the analysis here.

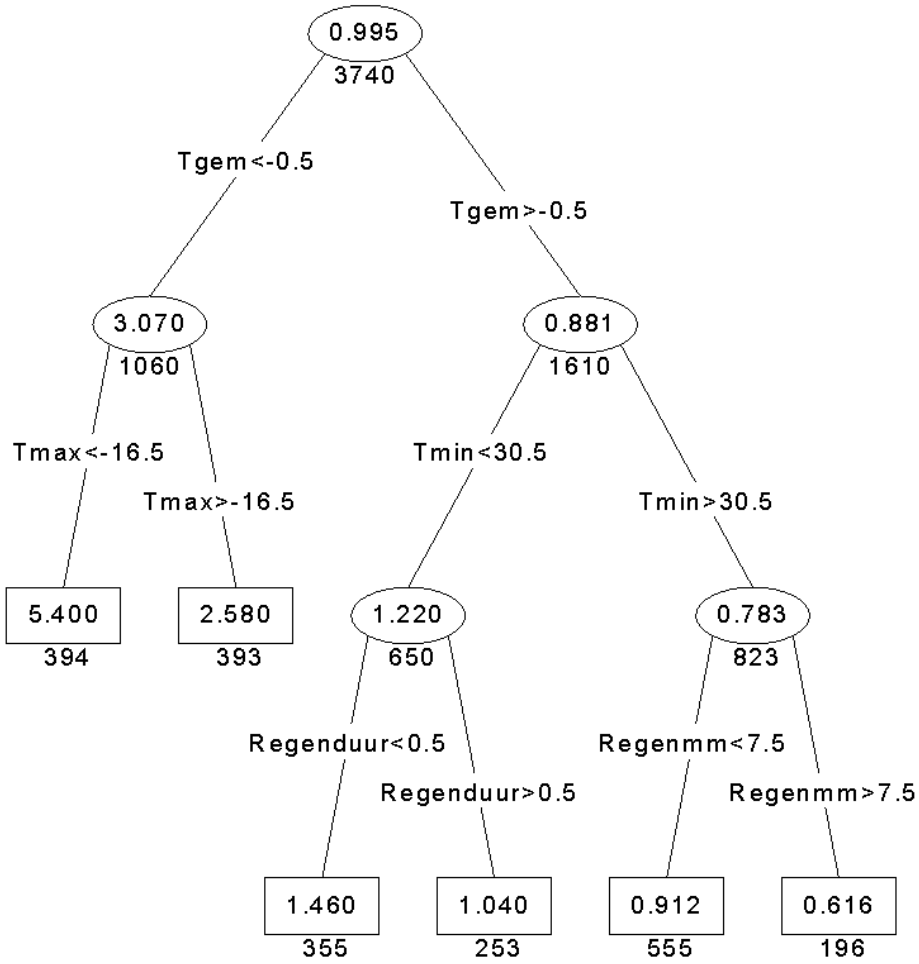

Figure 6 *Regression tree for SO_2 concentrations at Posterholt.*
Before estimating the tree, transformation (8b) was used to remove the trend from the data and to stabilize the variance over time. Transformed concentrations vary at around 1.0. Splitting variables for average daily temperature (Tgem), maximum daily temperature (Tmax) and minimum daily temperature (Tmin) are expressed in tenths of ˚C. Rainfall (Regenmm) is expressed in tenths of mm, and rainfall duration (Regenduur) in hours.

2.6 Validation and stability of trees

As well as providing a good fit within the relevant sample, our optimal pruned regression tree, and in fact any model, should give a good fit outside the data (compare criterion (e) in section 1.4). Therefore it is good practice to use the data available as a so-called *trainingset*, and to predict concentrations on other independent data; this is called *the testset*.

Criteria to evaluate the prediction performance of the tree to a *testset* or a *trainingset,* have been given in section 2.4. Now RT predictions on the *testset* data are generated by using the estimated tree, estimated in the *trainingset*, in combination with the explanatory variables available in the *testset*. In this way our predictions are not based on any information on concentrations falling in the *testset.* For the overall-mean predictor we use the mean for all concentrations in the *trainingset* as the predictor for the concentrations in the testset.

An example of cross validation is given in **Table 1**. Here, we evaluate the indices from equations (6) and (7) for:

- all daily SO_2 concentrations from Posterholt station (second column);
- all data minus the data of 1994 (third column);
- all concentrations occurring in 1994 (fourth column);
- all data minus the data for 2000 (fifth column);
- all concentrations occurring in 2000 (sixth column).

The table shows predictions during the *testset* periods 1994 and 2000 to be very good. Both indices $P_{\mu}^{*}(d)$ and $P_{naive}^{*}(d)$ in the *testsets* are higher than the corresponding values in the trainingset (compare percentages fourth row).

Table 1 *Indices from equations (5) and (6) for all data (1992-2001), all data except the days in 1994 (trainingset), all data in 1994 (corresponding testset), all data except the days in 2000 (trainingsset) and all data in 2000 (corresponding testset).*

Index	All data 1992 – 2001	All data, except 1994	Data 1994	All data, except 2000	Data 2000
$RE_{\mu}^{*}(d)$	0.57	0.57	0.20	0.56	0.31
$RE_{naive}^{*}(d)$	0.95	0.95	0.85	0.95	0.83
$P_{\mu}^{*}(d)$	43%	43%	80%	43%	69%
$P_{naive}^{*}(d)$	5%	5%	15%	5%	17%

The stability of an estimated tree could be low for two reasons. First, the tree may contain too many nodes. In this situation many explanatory variables are able to lower the overall deviance of the tree in the same small manner; furthermore, we can estimate a number of equivalent trees.

Second, a tree may be unstable if two or more explanatory variables are highly correlated. In Multiple Regression analysis this problem is called *multi-collinearity* and in the filter theory the *filter is said to be not observable.* Lack of observability means that the output of a filter is not uniquely determined by its input variables.

In Regression Tree analysis multicollinearity manifests itself in the choice of specific variable x_i to make a certain split. Reduction in deviance is reached for multiple x-variables if they are highly correlated. An example of multicollinearity in the examples throughout this report is shown in the variables:

- daily averaged temperature (variable Tgem);
- daily maximum temperature (variable Tmax);
- daily minimum temperature (variable Tmin), and to a lesser extent;
- global radiation.

Scatterplots for Tgem, Tmax and Tmin are given in **Figure 9**.

Are there solutions to the problem of multicollinearity? The answer is simply **no**. Statistics is not able to provide the exact unique relationships between some variables y_i, $x_{1,i}$ and $x_{2,i}$ if $x_{1,i}$ and $x_{2,i}$ are highly correlated. Information other than that based on statistical inferences should answer which variable should be coupled to y_i.

*In most cases the instability of Regression Trees is caused by multicollinearity, i.e. high correlation among the predictors **x**. To test for stability, we may compute the correlation matrix of all predictors. A cross validation on the final tree will reveal instability as well.*
Photo: H. Visser

We note that some researchers have applied *a principal component transformation* to all explanatory variables, making a new set of variables **x'** (principal components) that are uncorrelated with each other. The modelling is then performed between y_i and these principal components (e.g. Fritts, 1976). However, a serious drawback of this approach is that the selection-of-variables process (MR analysis) or the building of a tree (RT analysis) is performed on x_i' variables which have no clear physical interpretation. Therefore we do not advocate this approach.

2.7 Meteorological correction

We consider here two methods for calculating meteo-corrected concentrations on the basis of an optimally pruned tree. The first method has been proposed by Stoeckenius (1991). A meteo-corrected annual value or annual percentiles can be calculated with this method on the basis of frequency of leaves within a certain year relative to the frequency of the leaves for all years. The method of Stoeckenius has been described in detail in Dekkers and Noordijk (1997). In Noordijk and Visser (in prep.) refinements will be given to the method of Stoeckenius.

The second method has been described at the end of §1.3. The method gives meteo-corrected value for each day or month (this is not the case for the method of Stoeckenius). Annual averages or percentiles are simply calculated on these meteo-corrected days.

Throughout this report we will apply the latter method. An evaluation of both correction methods will be given in Noordijk and Visser (in prep.).

In **Figure 7** we have plotted the daily concentration for SO_2 at Posterholt, along with the RT predictions, based on the optimal pruned tree shown in **Figure 6**. The graph shows much overlap between measurements and meteo-corrected data, indicating that not much of the daily behaviour of SO_2 can be attributed to meteorological variations.

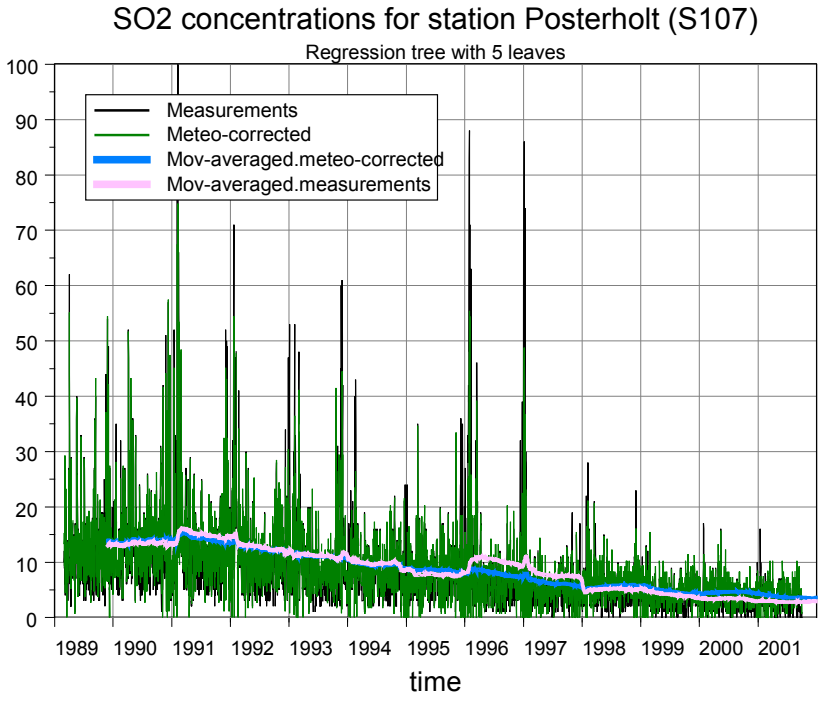

Figure 7 *Concentrations (black curve) and RT predictions (green curve) for daily SO_2 concentrations at Posterholt.*
The pink curve shows the moving average of concentrations using a window of 365 days, while the blue curve show the same for the meteo-corrected daily data.

2.8 Influence of sampling time

Finding rival models to a certain Regression Tree should not be limited to models based on other statistical principles. One should also check the sampling time of the measurements, y_i. If one analyses concentrations on the basis of *hourly data*, one will find mainly local meteorological conditions governing *daily variations.* If one uses daily averaged concentrations, one will find mainly meteo variables governing variation over the Netherlands within weeks. If one uses monthly averaged data, one will mainly find the meteo variables governing the annual cycle, on the scale of NW-Europe. Therefore it makes sense to check the sampling time when modelling concentrations.

As an example we have modelled the SO_2 concentrations for the Posterholt station using monthly averaged SO_2 concentrations and the transformation (8b). The Regression Tree is shown in **Figure 8A** and the corresponding time-series plot with concentrations and monthly RT predictions in **Figure 8B**.

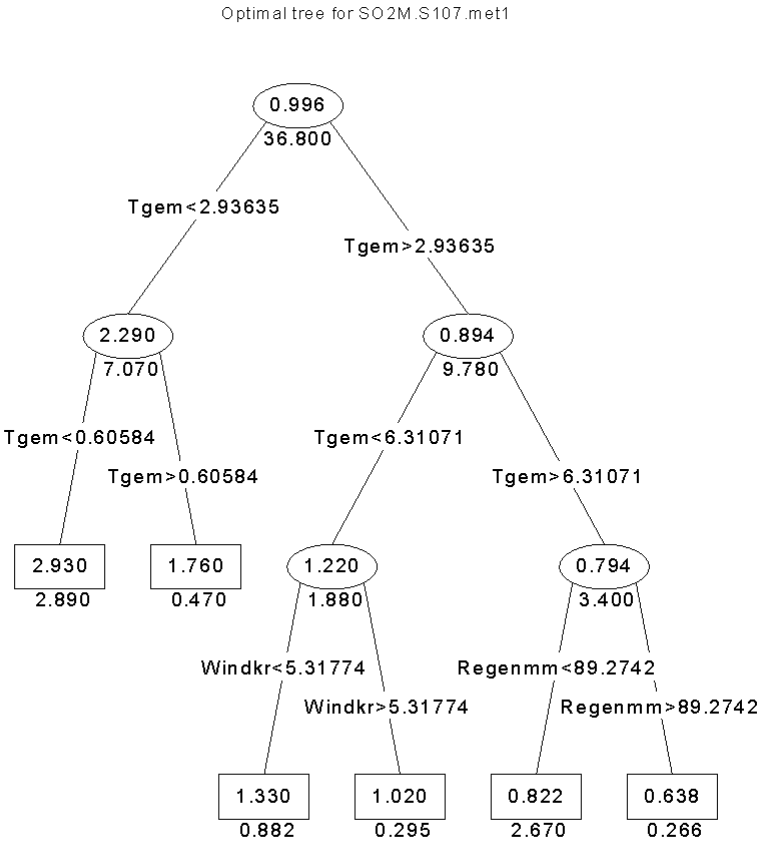

Figure 8A *Regression tree for monthly SO_2 concentrations at Posterholt.*
Before estimating the tree, transformation (4c) was used to remove the trend from the data and to stabilize the variance over time. Transformed concentrations vary at around 1.0. Temperatures (Tgem) are expressed in ˚C, total monthly precipitation (Regenmm) in mm, and wind speed (Windkr) in m/sec.

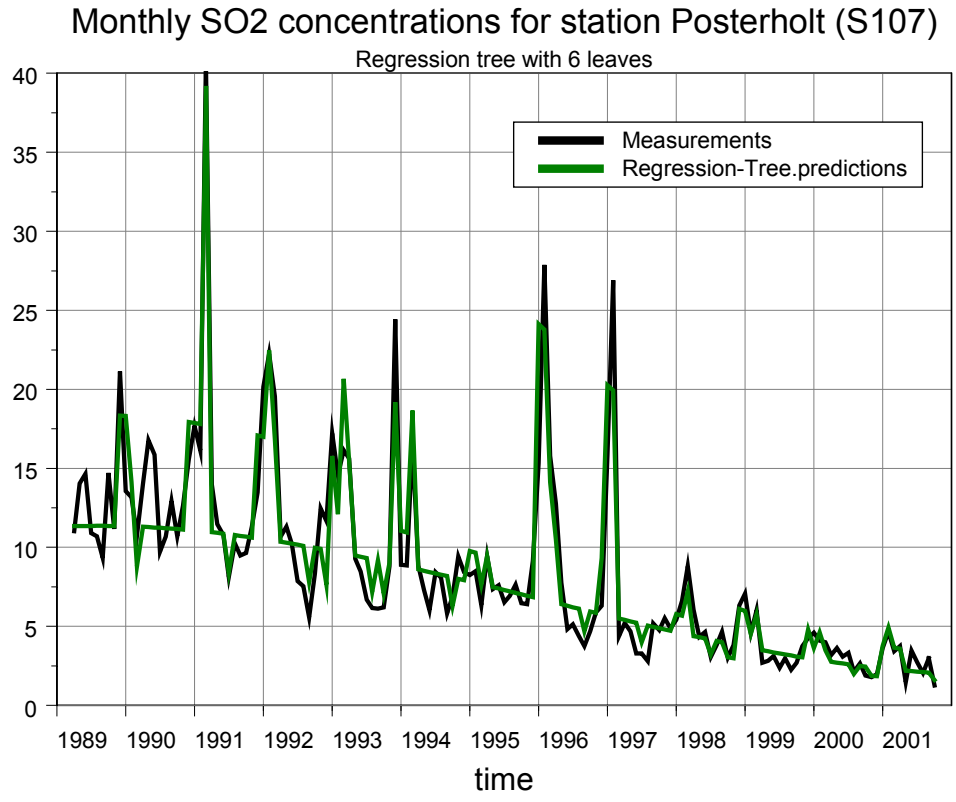

Figure 8B *Concentrations (black curve) and RT predictions (green curve) for monthly SO_2 concentrations at Posterholt. The model is based on the Regression Tree from Figure 8A.*

The difference in the RT based on daily and monthly data is clearly shown by the great differences in predictive power, as shown in **Table 2**. We can conclude from Table 2 that an RT based on monthly SO_2 data is much better in describing large-scale meteorological conditions similar to those reflected in inter-annual variation of concentrations. A similar result is found for PM_{10} concentrations (Chapter 5).

Table 2 *Indices from equations (5) and (6) for daily and monthly SO_2 concentrations. Data are for Posterholt station.*

Index	**Daily data 1992 – 2001**	**Monthly data 1992- 2001**
$RE_{\mu}^{*}(d)$	0.57	0.20
$RE_{naive}^{*}(d)$	0.95	0.17
$P_{\mu}^{*}(d)$	43%	80%
$P_{naive}^{*}(d)$	5%	83%

We note that by changing the sampling time from days to months, difficulties arise for wind direction. A monthly averaged wind direction has no meaning. For this reason we have added six dummy variables, each variable for one wind-direction sector of 60 degrees. Monthly averages of these variables now give the fraction of days in a month with wind direction in that particular wind sector. The sum of these monthly fractions is held to be 1.0.

2.9 Procedure in eight steps

In this document we aim to describe refinements of the approach of Dekkers and Noordijk (1997). Basically, the method proposed here consists of an eight-step procedure:

1. Check for outliers with descriptive statistics on the data. Explore the relationships between variables by scatter matrices.
2. Check to see if the original data should be transformed before RT analysis (e.g. to eliminate trends).
3. Classify the meteorological conditions. To this end an initial Regression Tree is estimated and then pruned to give the final Regression Tree.
4. Calculate leaf numbers, leaf predictions and meteo-corrected concentrations.
5. Validate the optimal Regression Tree by omitting certain years of data. By predicting these omitted data and comparing predictions with the real concentrations, an impression is gained of the predictive power and the stability (robustness) of the optimal Regression Tree (the principle of cross validation).
6. Make diagnostic checks on residuals and leaf frequencies for the optimal tree.
7. Visually inspect concentrations, predictions and meteo-corrected concentrations.
8. Compare variables in the optimal Regression Tree as well the predictive power of the tree with those found by estimating a Multiple Regression model on exactly the same data.

As a final step we may fit a low-pass filter through the meteo-corrected concentrations, which allows the estimate of 90% confidence limits. In general, the fit is based on a second-order polynomial (a 'parabola fit'). The filter separates an emission-related concentration trend and noise. However, it assumes consistency in the measurements and a smooth change in emissions (no sudden changes). Because of this assumption we name this step ‘optional’.

2.10 Limitations

The Regression Tree approach also has limitations. We recall two drawbacks here of importance for the interpretation of the Regression Tree results (compare §2.6):

- highly correlated predictors induce instability in tree estimates;
- if both response variable y_t and selected predictors $x_{i,t}$ contain long-term trends, the meteo-correction procedure could become sub-optimal. The method cannot uniquely distinguish between emission-induced changes in concentration and meteorologically-induced changes.

We provided guidelines in §2.6 to detect potential instability of tree estimates by the reasons mentioned above. However, if such instabilities occur, results should be presented with care.

3. ART software

3.1 S-PLUS functions and scripts

A state-of–the-art software library on CART analysis is available within S-PLUS. A general description on S-PLUS for RIVM is given by Dekkers (2001). The implementation of Regression Tree routines is described in S-PLUS (2000) and Venables and Ripley (1997).

We have extended the RT software library of S-PLUS with a number of scripts and functions to enable the analyses and tests described in the preceding chapter. We have named this software tool ART (*Air pollution by Regression Trees)*.

ART consists of a number of S-PLUS scripts and routines. In the S-PLUS script (**Appendix A**) a number of these functions are named. Details on these functions and additional scripts will be given in Visser (in prep.). In **Chapter 4** we will give a step-by-step example of the use of ART based on the script given in **Appendix A**.

The Regression Tree methodology is implemented in ART (Air pollution and Regression Trees). This software is based on S-PLUS routines. Photo: H. Visser

3.2 Preparing data prior to analysis

Because the ART software is implemented within S-PLUS, one has to import air pollution data and meteorology into S-PLUS. Data structures in S-PLUS are called *dataframes*. In Visser (in prep.) a detailed description is given on how to import air pollution data from a number of stations into one data frame. A script is also given for the import of meteorology and divided into five regions over the Netherlands, leading to the generation of 5 data frames.

As a final step, all air pollution stations are coupled to the meteo-region data frame to which they correspond, leading to 5 additional data frames. ART analyses are performed on these final data frames.

4. Regression Tree analysis step by step

In this chapter we will demonstrate the estimation of a proper Regression Tree with corresponding concentration predictions and meteo correction as a 7-step procedure. For illustrative purposes we have made a simulated example (**Appendix G** in Visser, 2002; the yellow area).

Here, a data frame, **Pseudo5,** is generated with a dependent variable y_t = Tgem/10, i.e. our air pollution component is 100% linear coupled with air temperature (divided by 10 to obtain temperature in ^oC). Explanatory variables are:

1.	Tgem	daily averaged temperature in tenths ºC
2.	Tmin	minimum hourly temperature in tenths ºC
3.	Tmax	maximum hourly temperature in tenths ºC
4.	RHgem	relative humidity in %
5.	Regenmm	amount of precipitation in tenths mm
6.	Pgem	air pressure in mbar
7.	Stralgem	radiation in J
8.	PercStral	percentage radiation
9.	Windr	wind direction in degrees
10.	Windkr	wind speed in m/sec
11.	Pasqudag	Pasquill stability class for daytime conditions
12.	Pasqunacht	Pasquill stability class for night-time conditions

This example is also interesting because we can see how RTA handles (perfect) linear relations between y_t, and one or more explanatory variables. This is typically a situation where Multiple Regression (MR) will find the right relationship because MR estimates linear relations by definition.

Furthermore, we know the perfect prediction for each day, $Tgem_t/10$, *in advance*. We also know the meteo-corrected daily value of 'Index' in advance. Because all variations in the series 'Index' are due to meteorological factors, the meteo-corrected 'Index' should be constant over time. This value is the average value of the variable Tgem/10 over all 3653 days in the period 1992-2001: 10.3 ºC. How well will our Regression Tree estimation procedure reconstruct these values?

The eight-step procedure is given in Appendix A as an S-PLUS script. In the following sections we will describe these steps using the data from data frame **Pseudo5**.

4.1 Step 1: looking at the data

We block **aa.dat <- Pseudo5[-1:-365,]** (yellow area in Appendix A) in the SPLUS script **RTAonPseudoseries**. In doing this we copy our dataframe, **Pseudo5** to the general dataframe **aa.dat**. The addition [-1:-365,] means omitting the first 365 days in the dataframe (this is the year 1991). Now the analysis covers the years 1992 through 2001, where N = 3653 days.

Second, we block **m<- ….** , to generate general statistics for each of the variables in our dataframe. This is for a general check on the input data. How many data are missing and are there specific outliers?

Third, we block **guiplot…** to generate a scatter-plot matrix between all variables involved. The first scatter-plot matrix for **Pseudo5** is given in **Figure 9**. Here, we see the perfect linear relation between variable Index (our dependent variable y_t) and explanatory variable Tgem. y_t is also highly correlated to variables $Tmin_t$ and $Tmax_t$, which reflects the high correlation between daily averaged, daily maximum and daily minimum temperatures.

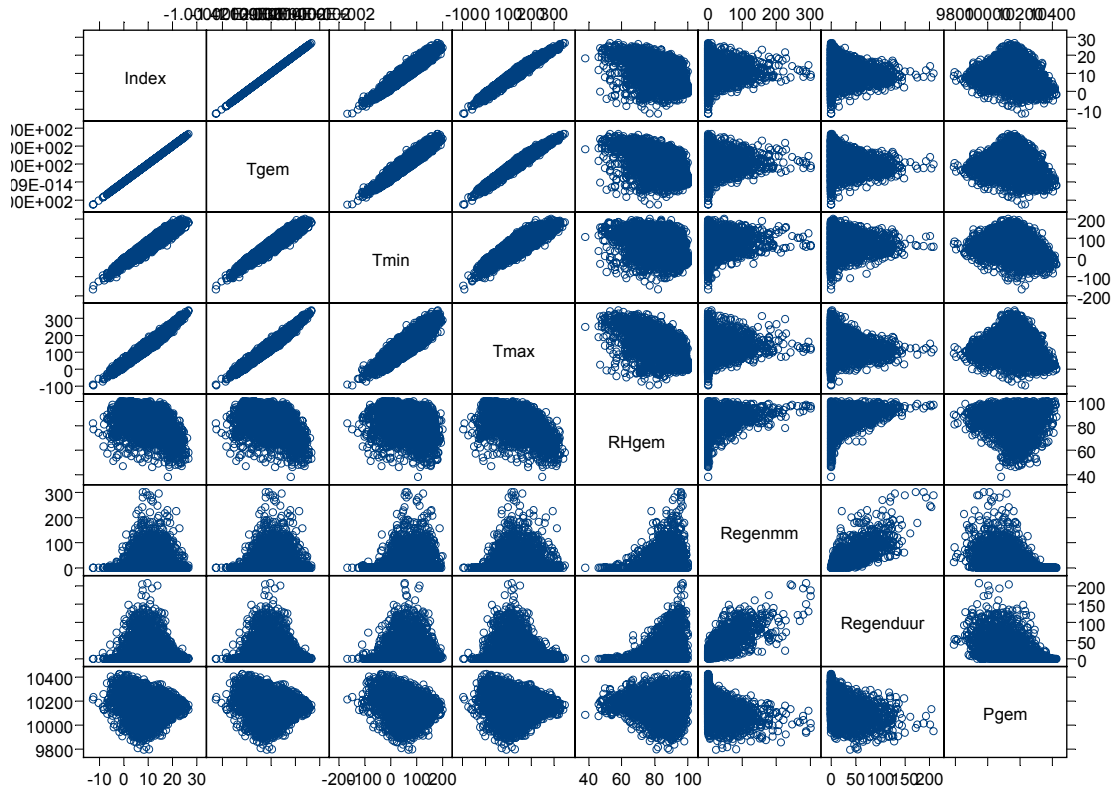

Figure 9 *Scatter-plot matrix for eight variables from **Pseudo5**. Variable 'Index' is the dependent variable. Time t is in days and runs from 1992 to 2001 (N= 3653).*

Finally, we complete the text string vector **luvo** with the variable names of the 12 explanatory variables mentioned above. We note here that the string vector **luvo** may contain any subset from the 12 explanatory variables.

4.2 Step 2: transformation of concentrations

Step 2 is marked in the colour green in Appendix A. In this step we verify if data should be transformed before performing a Regression Tree analysis. See §2.6 for details. **Figure 10** shows the range-mean plot. The points lie more or less on a horizontal line, indicating that the variance for each year does not depend on the corresponding annual mean concentration.

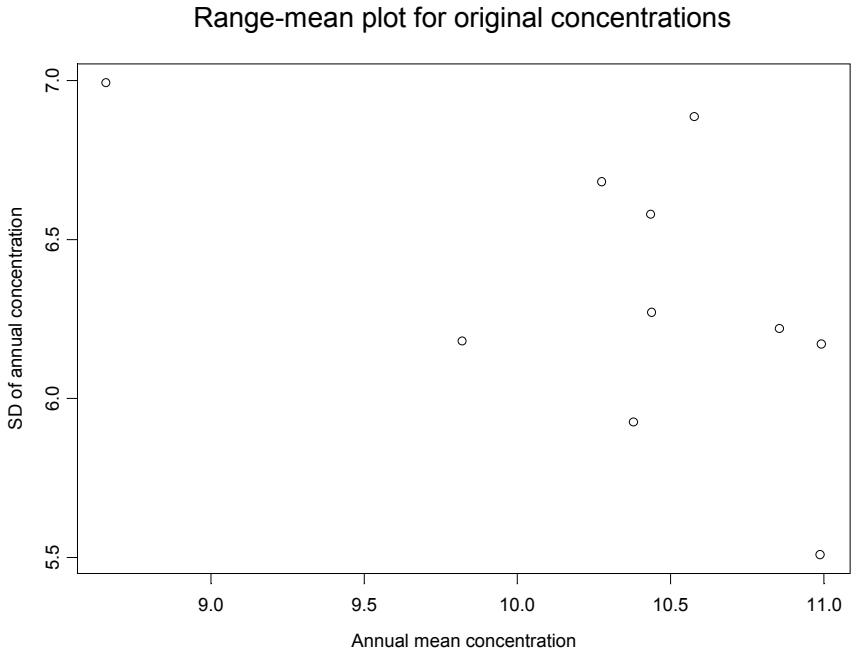

*Figure 10 Range-mean plot for variable **Pseudo5**.*

A time-series plot with the long-term trend is given in **Figure 11**. The graph confirms our findings from **Figure 10**: no transformation is needed for this example.

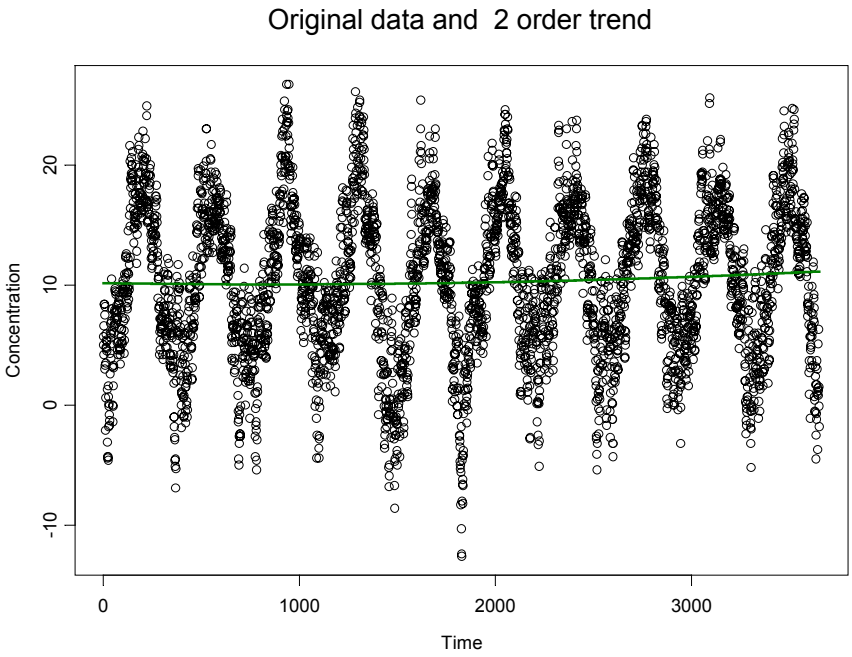

*Figure 11 Time-series plot for **Pseudo5** with estimated trend (green line).*

4.3 Step 3: initial Regression Tree and pruning

Step 3 is marked in the colour blue in Appendix A. In this step we block **f.RTAnew (aa.dat, "Pseudo5", "Index", luvo, 1, 80, 40)**. The second argument, Pseudo5, is used as a unique identification of files and graphs. The same holds for Index, which, at the same time, is the name of the y_t variable. The raw initial tree is given in **Figure 12**. This graph is automatically displayed by S-PLUS.

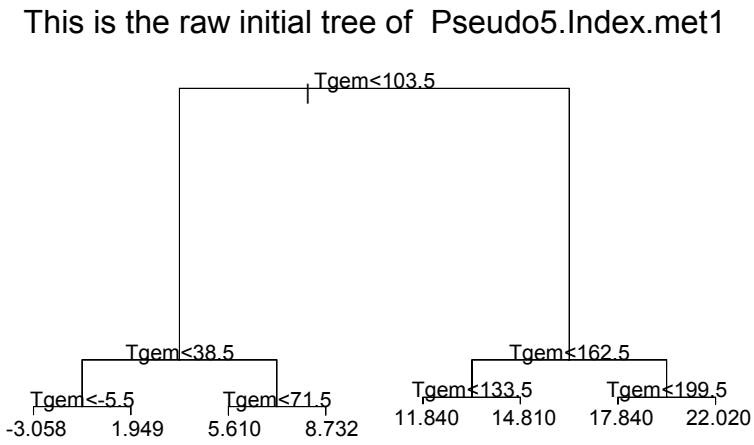

*Figure 12 Raw initial tree with identification **Pseudo5.Index.met1**.*

From **Figure 12** we see that the trees contains eight leaves and all splits are done using variable Tgem. Note that the variable Tgem has a unity of tenths of ºC, so the values should be divided by 10 to obtain the value in ºC.

Second, we obtain a menu for choosing a lower number of leaves, used for pruning of the initial tree. We can decide to prune on the basis of the cross-validation plot given in **Figure 13,** which is also automatically displayed by S-PLUS. We see from **Figure 3** that the number of eight leaves is optimal and, therefore, we decide to keep the full tree with eight leaves (in real applications we will choose a number much lower than the maximum number of leaves found in the initial raw tree).

The final tree, which in our case is equal to the raw initial tree, is saved in a Postscript file **c:RTAdataOri/Pseudo5.Index.met1**. By clicking this file, **GSview** will display the final tree in a format more elegant than that shown in **Figure 12**. The tree can be blocked by choosing **Edit** and **Copy C** in **GSview**. Then it can be displayed in a Word file by typing **Cntrl V**. See **Figure 14** for this final tree.

All calculated results are sent to the screen and automatically printed on the standard black-and-white printer.

If a transformation is chosen (variable **Trans** in the argument list of function **f.RTAnew** set to '2' or '3'), S-PLUS will generate plots of the original data with the estimated trend, as well as a plot of the de-trended data. Now the Regression Trees will be estimated on these de-trended concentrations. If the estimated trend is too inflexible, a more flexible trend may be estimated by setting the last argument from **f.RTAnew** to a higher number (a polynomial with a higher order is chosen).

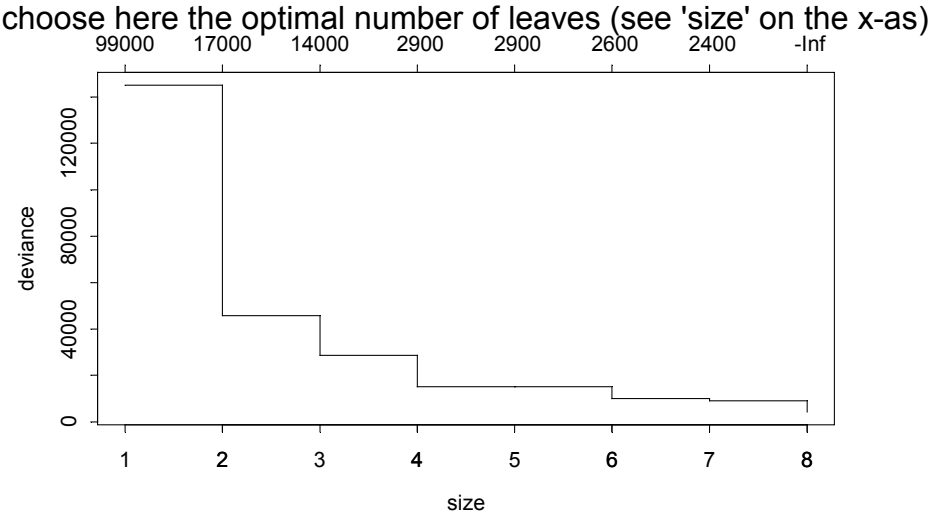

Figure 13 *Cross-validation plot for finding the optimal number of leaves (plotted on the x-axis). The curve has no clear minimum. Therefore, we choose eight as the optimal number of leaves.*

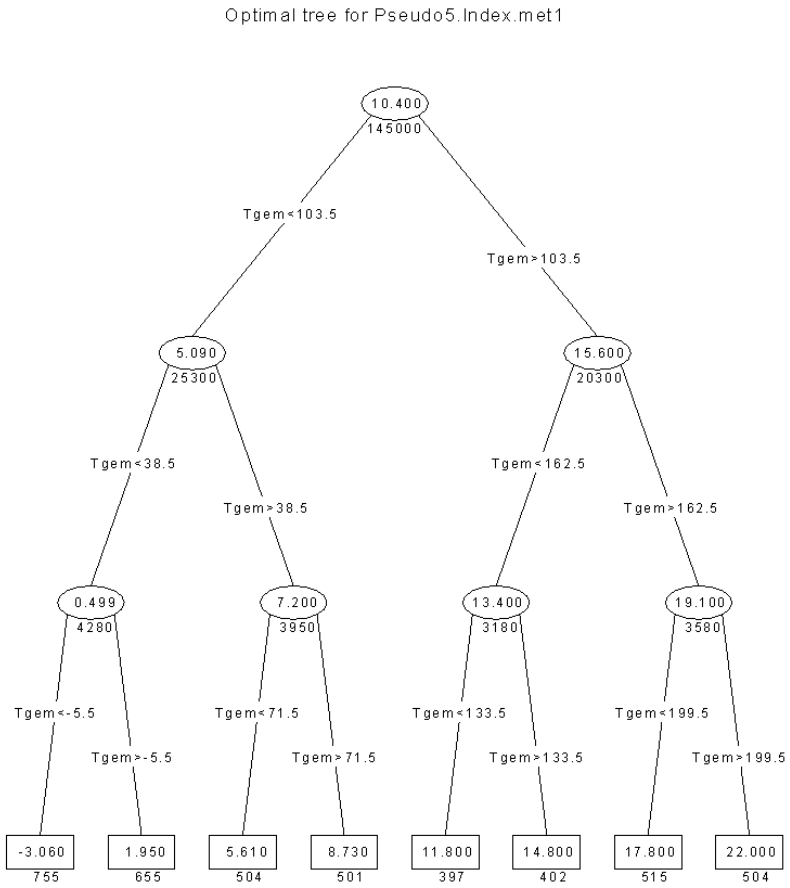

Figure 14 *Postscript version of the final tree for data frame **Pseudo5**.*
Note that variable Tgem is in tenths of ºC. The ellipses in the tree are the nodes, and the rectangular boxes are the final nodes or 'leaves' of the tree. The values in the ellipses and boxes are the averages of the concentrations of all days falling in that specific node. These values are also used as predictions (if the original data have not been transformed). The figure beneath each box stands for the deviance of the data belonging to the node, i.e. the sum of the squared deviances $(y_i - y_{mean})^2$ for all i belonging to a specific node.

4.4 Step 4: meteo-corrected annual data

Step 4 is marked in **Appendix A** by the colour 'white'. The function **f.RTAfreq** calculates the daily predictions and meteo-corrected daily data. These data are sent to the file **c:RTAdataOri/Pseudo.Index.met1.dat**. Calculated statistics are given on the screen and printed automatically to the standard black-and-white printer.

We will mention two important results from the output. First, on prediction errors (§2.7): the output shows (i) the mean squared prediction errors (R*(d)), (ii) idem for a model which has the constant prediction ymean for all days (R_μ*), and (iii) idem for the "naive model" (R_{naive}*). In the latter model we use for the prediction of the value y_t simply the value y_{t-1} of the preceding day. From the output we read that the Regression Tree predictions are 97% better than the predictions by the "ymean model" (P_μ*(d)) and 71% better than the "naive model" (P_{naive}*(d)).

Second, the output shows that meteo-corrected concentrations vary between 10.3 and 10.5. This holds for both meteo-correction methods. The real meteo-corrected value is 10.3 for all days. Thus, the regression tree with eight leaves is very well able to reconstruct the right value.

S-PLUS automatically generates five plots. The first plot is given in **Figure 15**. It is a good illustration of how a Regression Tree approach deals with linear relations between y_t and one or more explanatory variables. We have 8 leaves and therefore 8 values to choose for a prediction for a specific day. The Regression Tree has simply split the temperature variable 'Tgem' into 8 non-overlapping intervals and uses the average of the temperatures within a leaf as the prediction of a specific day.

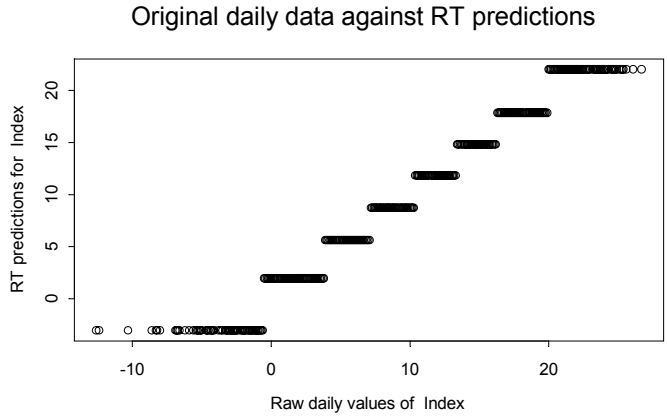

Figure 15 Scatterplot for daily concentrations (x-axis) and Regression Tree predictions for corresponding days.

The meteo-corrected annual values for y_t (the variable 'Index') is given in **Figure 16**. The Figure shows the correction procedure according to Stoeckenius (lower graph) and the method described in §2.4 (upper graph). As mentioned above, the Regression Tree approximation is good, but slightly above the real value of 10.3 (annual predicted values vary between 10.3 and 10.5). We also note that both meteo-correction methods yield identical annual corrected Indices (**Figure 16**).

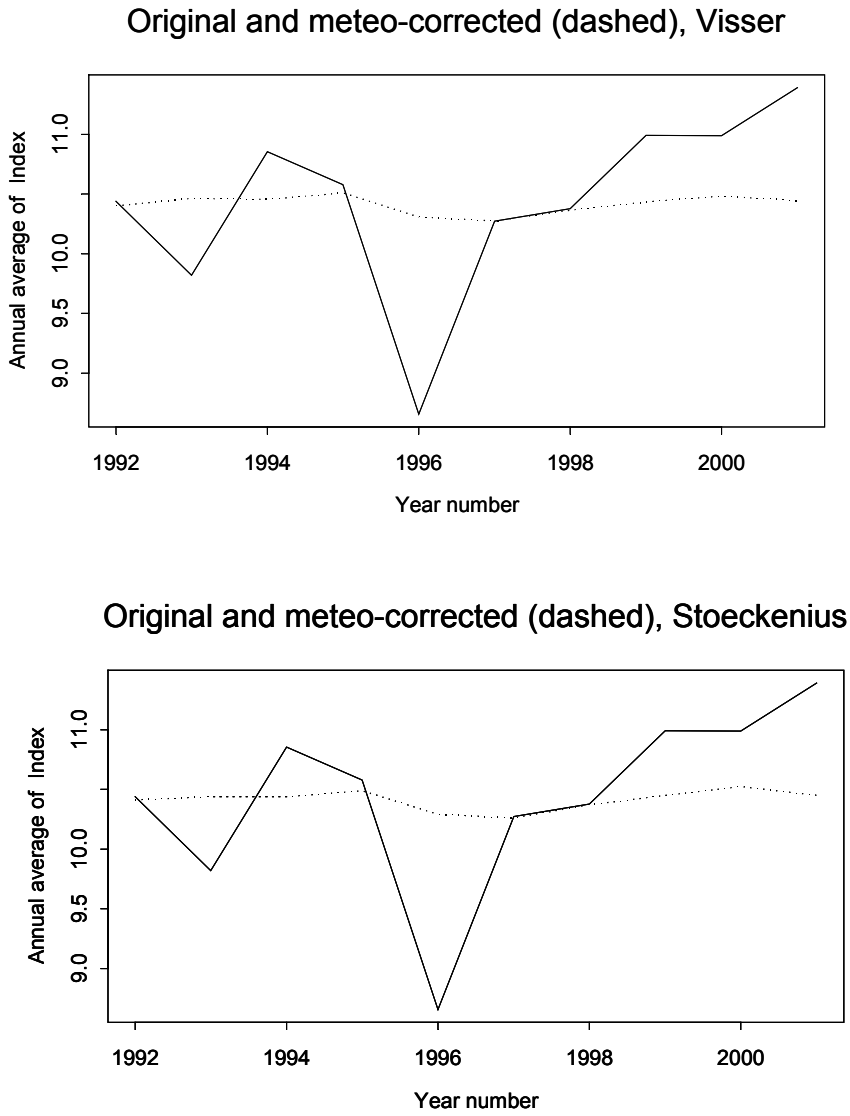

Figure 16 *Annual averaged concentrations with meteo-corrected annual values, based on the correction procedure of Stoeckenius (upper graph) and the correction method described in §2.4 (lower graph).*

4.5 Step 5: validation of the final Regression Tree

Step 5 is marked in **Appendix A** by the colour light grey. In this step we estimate the optimal tree for all years except 1995 and 1999. The daily values for the years 1995 and 1999 are then predicted using the estimated tree with eight leaves. In the output we can read how the predictions are in relation to the 'ymean model' and the 'naive model'. We can also compare these results with those found in the preceding section.

The result is that the predictions for the years 1995 and 1999 are 97% better than 'ymean' (P_{μ}*(d)) and 75% better than 'naive' (P_{naive}*(d)). Estimates over all the years except 1995 and 1999, i.e. the years for which we have estimated the Regression Tree, yield 97% and 71%, respectively (in step 4 we found identical percentages). This result shows that our regression tree approach is very stable if data are left out (in this example, we have left out 20% of the data).

4.6 Step 6: diagnostic checks

Step 6 is marked in **Appendix A** by the colour pinkish grey. Three diagnostic graphs are generated in this step. The first graph is shown in **Figure 17A**. Here, we see that for each year the frequency of the eight leaves is expressed in days on the x-axis (each panel representing one of the years, 1992 through 2001).

The second plot is given in **Figure 17B**. Here, we see that for each leaf the frequency of the 10 years (1992 – 2001) is expressed in days on the x-axis (each panel representing one of the eight leaves). First, this plot is important for finding effects of meteorological changes over a range of years: for example, the bar chart for leaf 1 shows a decreasing tendency over 1992-2001. As the method corrects for meteorological fluctuations and for systematic drifts in meteorology (e.g. climatic change), the trend in meteo-corrected concentrations may differ from the trend in the original concentrations. Second, the plot can be used for the detection of missing meteo classes in a specific year.

The third plot is given in **Figure 17C** and shows for each leaf a histogram for each leaf of all y_i values falling in that specific leaf. A number of skewed histograms may indicate that the original y_t values should be transformed (the splitting of branches in the tree is based on variances, but the variance of skewed data is less meaningful).

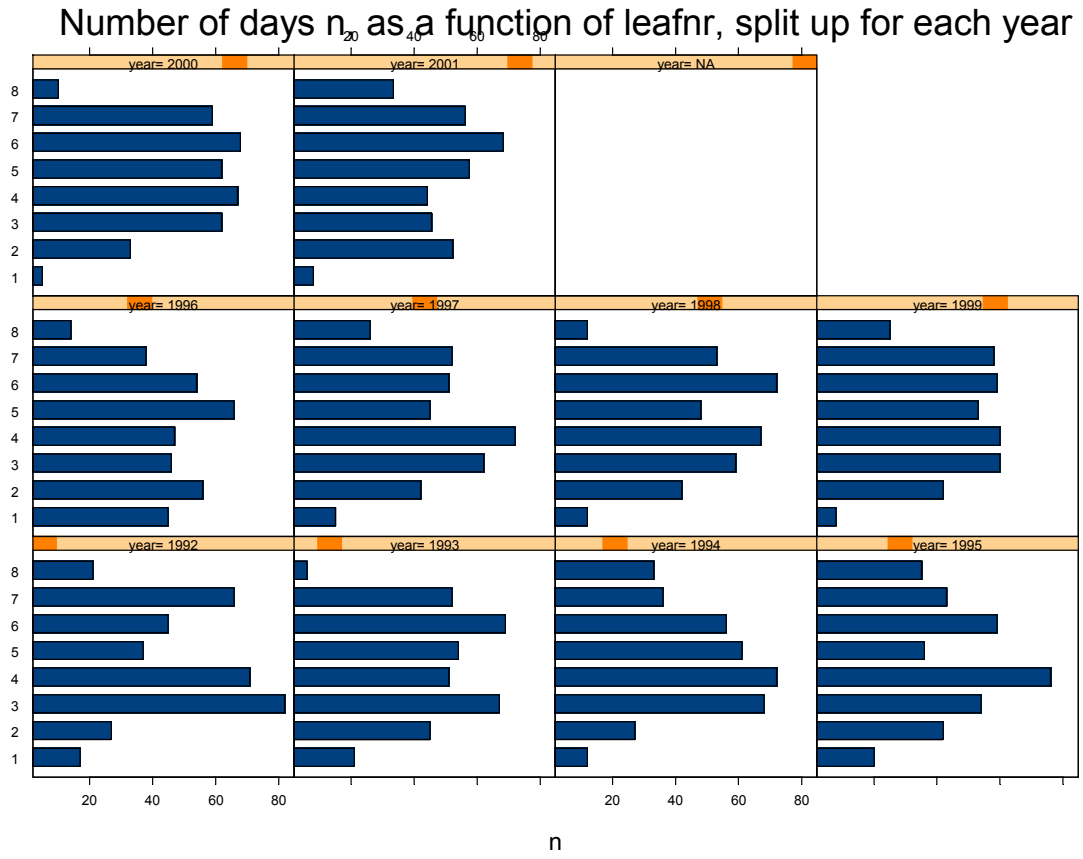

Figure 17A *Trellis plot with histograms for each year, with each histogram showing the number of the leaves (1 through 8) on the y-axis and the number of days on the x-axis in that specific year, with concentrations falling in that specific leaf.*

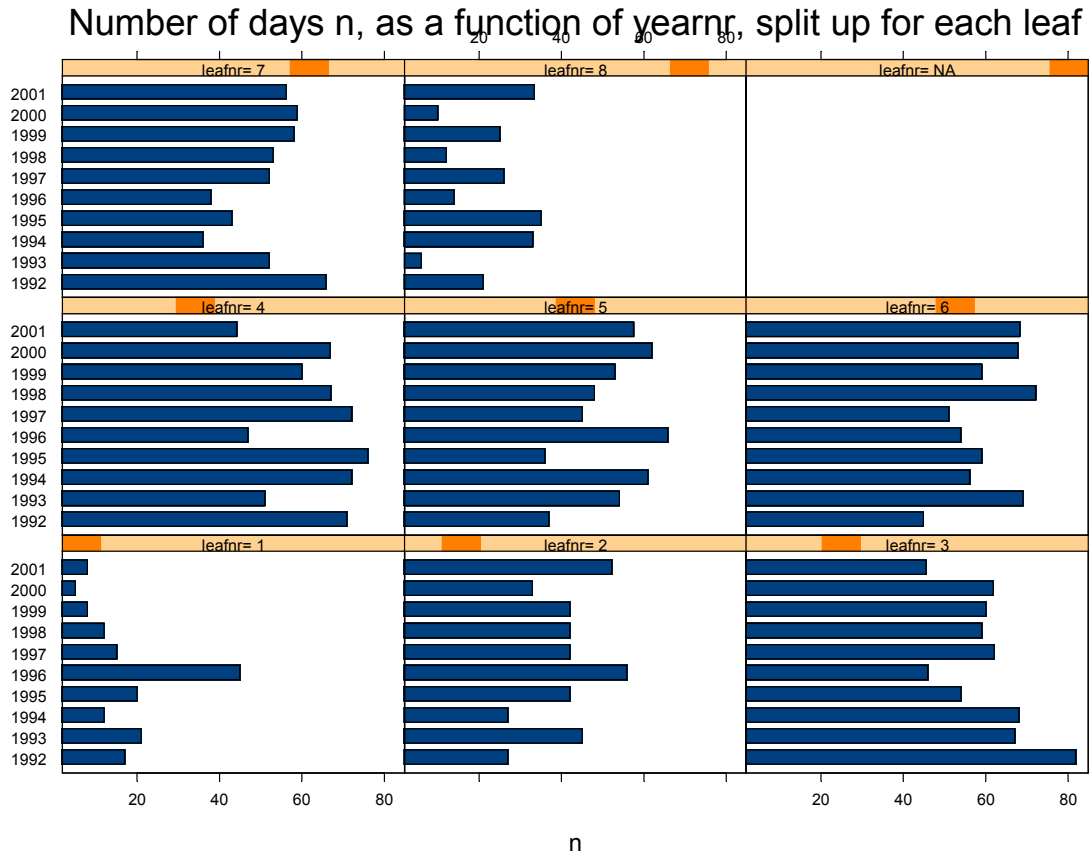

Figure 17B *Trellis plot with histograms for each leaf, with each histogram showing the number of years (1992 through 2001) on the y-axis and on the number of days on the x-axis for that specific leaf, with concentrations falling in that specific year.*

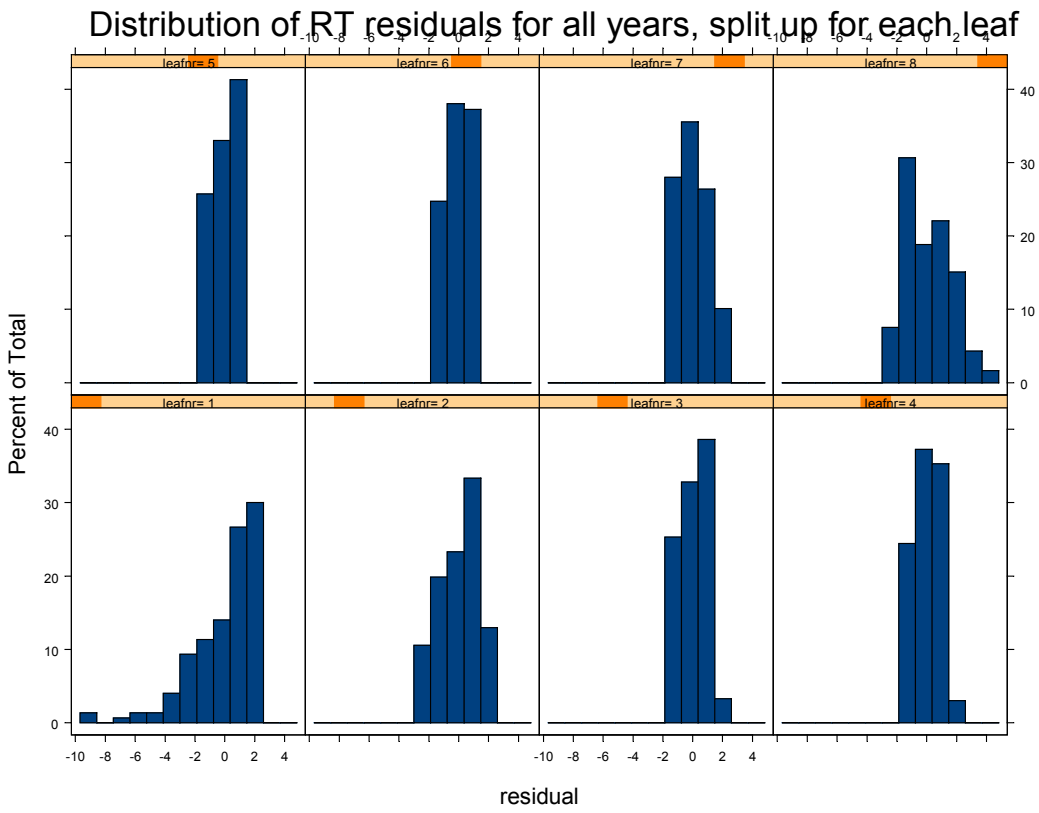

Figure 17C Trellis plot with y_i histograms for each of the 8 leaves, index i covering all days with concentrations for that specific leaf.

4.7 Step 7: visual presentation

Step 7 is marked in Appendix A by the colour red, where we made a plot using the script **RTAplotPredictions**. The result is shown in **Figure 18**. The blue curve is very close to the *real* meteo-corrected value of 10.3.

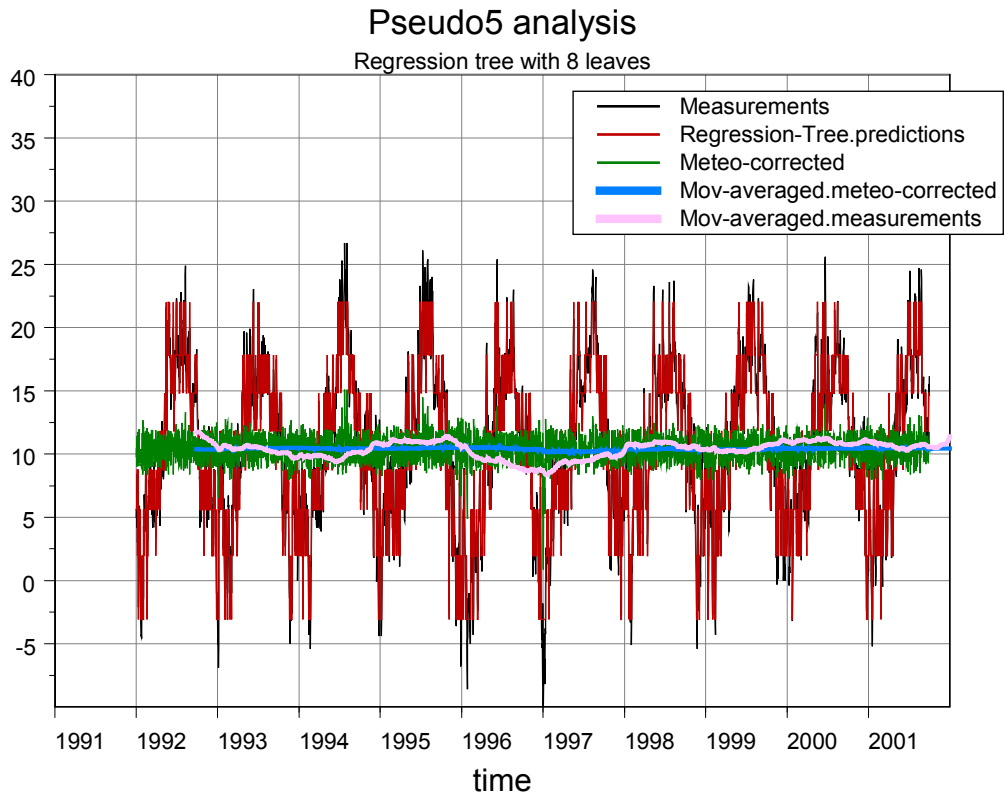

Figure 18 *Plot of y_t variable Index (red curve) and the daily RT predictions (green curve).*
The black curve represents the moving average of the Index using a window of 365 days (if less than 270 days are available, the value NA is returned). Analogous to this, the blue curve represents the moving average of the daily meteo-corrected Index values. Values of the black and blue curve at the vertical grid lines are equal to the annual values shown in Figure 16.

4.8 Step 8: Multiple Regression versus Regression Trees

Step 8 is marked in Appendix A by the colour dark green. A Multiple Regression analysis was performed on y_t and all 12 explanatory variables. As was expected, MR detects 'Tgem' as the one and only variable with a perfect fit: $R^2 = 1.0$ and all residuals zero!

5. PM_{10} at nine regional stations

As a case study we have chosen nine regional PM_{10} stations in the RIVM National Air Quality Monitoring Network (LML), which were in operation over the years 1992-2001. We have left out station Witteveen in this case study. Concentrations for this station show a decreasing trend due to environmental changes at the monitoring site. For general information on PM_{10} we refer to Visser et al. (2001), and Buringh and Opperhuizen (2002a,b).

The location of the stations is shown in **Figure 19**, with the caption also naming the division into 3 groups: regional stations, city stations and street stations. In this Chapter we will give the results for the nine regional stations only. The meteorological information has been provided by KNMI and contains the variables listed at the beginning of Chapter 4.

5.1 Analysis at individual stations

We started by performing analyses on the daily averaged PM_{10} levels. This, however, resulted in only 35% of explained variance in the PM data. For the second analysis, based on the monthly average PM values, the meteorologically explained variance rose to 60-70%. This remarkable difference may partly be explained by the phenomenology of PM. PM stays in the ambient air for a couple of days until it is removed by either dry or wet deposition. Due to the wind regime in the Netherlands, the ambient concentrations are influenced to a large extent by the situation in neighbouring countries. The meteorological factors used here, however, only pertain to the meteorology for the Netherlands; its variance on a daily basis does not say very much about the situation in the rest of Western Europe. Monthly meteorological averages in the Netherlands are more influenced by large-scale meteorology and form, therefore, a better basis for meteo correction of PM.

The results show that the main meteorological factors explaining the variance are rainfall, temperature, wind speed and wind direction. Continental wind directions result in higher concentrations, just as days with sub-zero temperatures. At the coastal stations temperature seems to be the main variable; in the east and south of the Netherlands the rainfall is a slightly stronger explanatory variable for local concentrations of PM. The lowest monthly averages varied from 20-30 $\mu g/m^3$ and the highest reached values of 60-70 $\mu g/m^3$.

The year 1992 is less reliable, because a number of the time-series in this first year of measurements are incomplete and because of some starting problems in the measurement technique.

An example is Vredepeel (code 131), shown in **Figure 20** . This figure shows the estimates for monthly data and the estimated tree. Main splits are on temperature (variable Tgem), precipitation (variable Regenmm) and wind speed (variable Windkr). Because of the trend transformation (equation (8b)), node values vary around 1.0.

Figure 19 *Map of the Netherlands giving the PM_{10} monitoring locations. Codes are explained below. Stations in green are regional; stations in blue are city background stations and stations in red are city street stations.*

131	Vredepeel-Vredeweg
133	Wijnandsrade-Opfergelstraat
230	Biest Houtakker-Biestsestraat
318	Philippine-Stelleweg
437	Westmaas-Groeneweg
444	De Zilk-Vogelaarsdreef
538	Wieringerwerf-Medemblikkerweg
722	Eibergen-Lintveldseweg
724	Wageningen-Binnenhaven
929	Witteveen-Talmaweg (omitted from analysis)
404	Den Haag-Rebecquestraat
418	Rotterdam-Schiedamsevest
441	Dordrecht-Frisostraat
520	Amsterdam-Florapark
236	Eindhoven-Genovevalaan
433	Vlaardingen-Floreslaan
639	Utrecht-Erzeijstraat
641	Breukelen-Snelweg
728	Apeldoorn-Stationsstraat

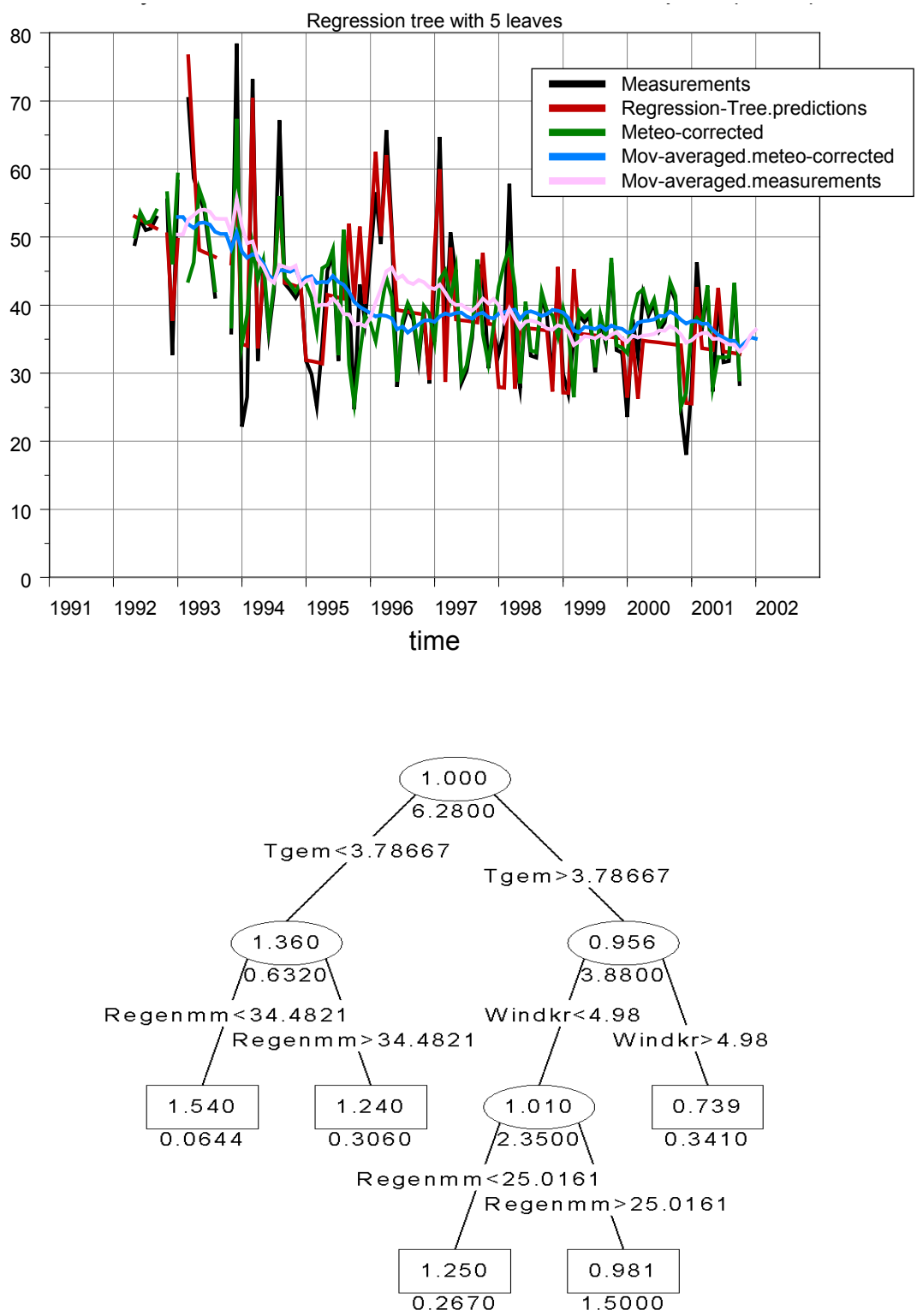

Figure 20 *Regression Tree graphs for the regional PM_{10} station, Vredepeel.* The upper graph shows the monthly averaged concentrations in µg/m^3, the predictions, and the meteo-corrected concentrations. Also given are the 12-month moving averages of the original concentrations (pink curve) and the meteo-corrected concentrations (blue curve). The lower graph shows the corresponding tree with five leaves. Splitting variables are temperature (Tgem) expressed in ˚C, precipitation (Regenmm), in mm and wind speed (Windkr), in m/sec.

Highest PM_{10} concentrations occur if monthly averaged temperatures are below 3.8 ºC and *at the same time* monthly precipitation totals are below 34 mm. *Lowest* concentrations occur if monthly temperatures are above 3.8 ºC and at the same time monthly averaged wind speed is above 5.0 m/sec.

From the upper graph in Figure 20 we see that the moving averages of both concentrations and meteo-corrected concentrations decrease similarly from 52 $\mu g/m^3$ in 1992 to 36 $\mu g/m^3$ in 2001, i.e. a decrease of 16 $\mu g/m^3$ in a ten-year period! Furthermore, concentrations are elevated in 1996 due to two unusual cold, dry winter months.

From these results we can conclude that the long-term downward trend is **not** influenced by meteorological variability.

Figure 21 confirms this conclusion from a slightly different angle. This Trellis graph shows the frequency from 1992 for each leaf (= meteo class) of Figure 20. The numbering of leaves (1-5) corresponds to the number in the lower graph of Figure 20 (simply follow the leaves from left to right). *None* of the five panels shows a clear trend over time. This result is consistent with the moving-average trends in Figure 20.

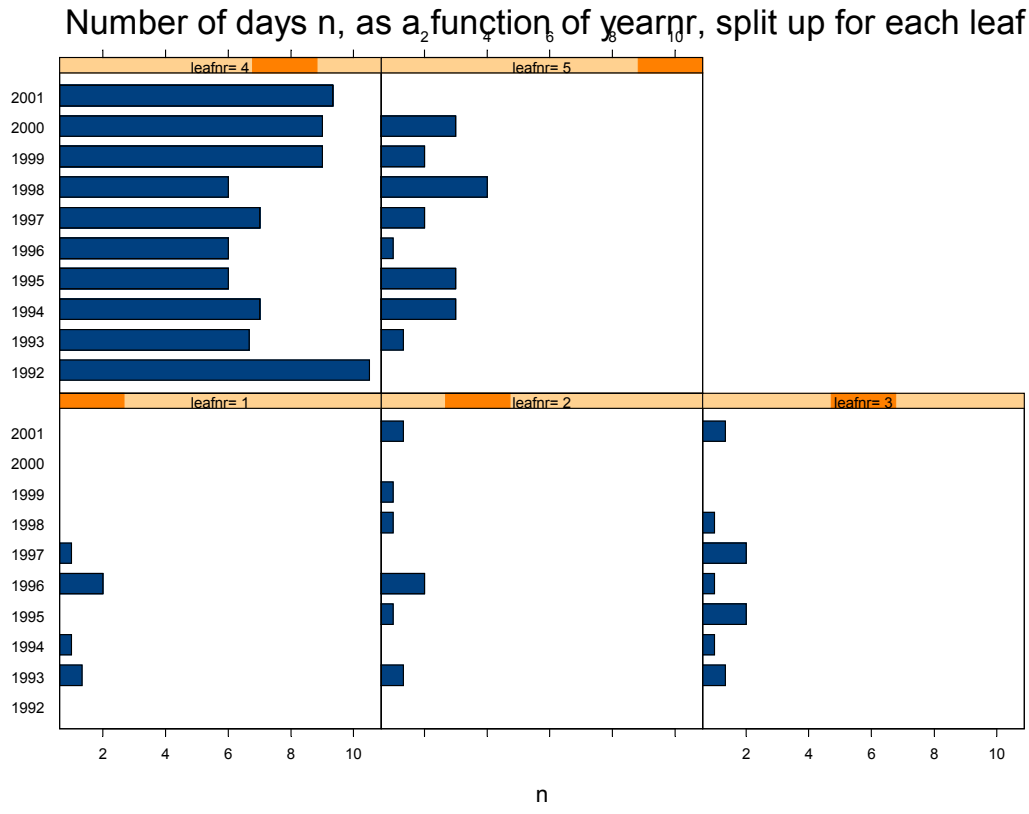

Figure 21 *A trellis graph showing the frequency of occurrence, expressed as the number of months (x-axis), as a function of year (y-axis).*
Each panel corresponds to one of the five leaves shown in Figure 20 (leaf #1 is at the far left of the tree).

5.2 Regional averages

Results for stations other than Vredepeel were similar, although Regression Trees differed on details. Therefore we have annual averaged concentrations and meteo-corrected concentrations for all nine regional stations, leading to two regionally averaged curves. These curves are presented in **Figure 22**.

The annual regional concentrations clearly show a downward trend over the years. The annual averaged measurements of all nine regional stations (red curve in **Figure 22**) decrease from 43.3 µg/m^3 in 1992 to 31.7 µg/m^3 in 2001. In other words, a decrease of 11.6 µg/m^3 occurs in a ten-year period; relative to the year 1992, this decrease is 27%. As mentioned in the preceding section, the year 1992 was the starting year of the PM_{10} monitoring network. We have estimated that concentrations in this specific year were probably ~1 µg/m^3 too high. Therefore the decrease is slightly lower: **~25%.**

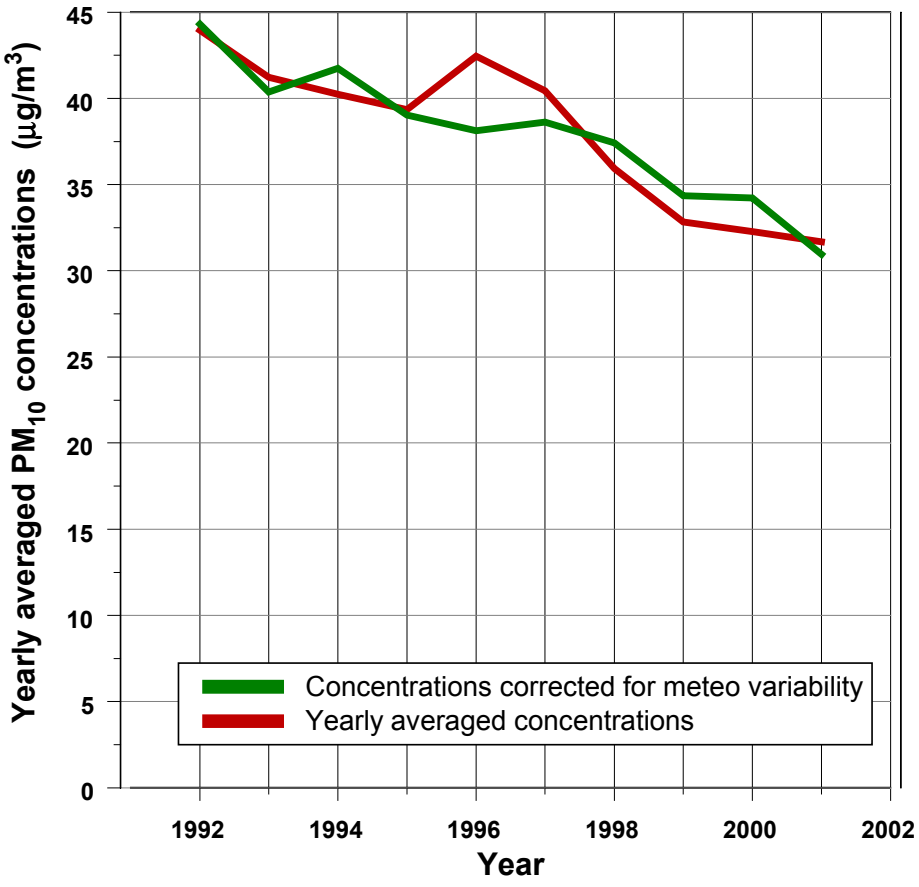

Figure 22 *Annual-averaged concentrations for nine regional PM_{10} stations in the Netherlands.*

After correction for meteorology there seems to be difference in the PM concentrations for the period 1992-1998 and the post-1998 period. Up to 1998 the decrease in PM levels seems limited. Afterwards, the decrease in concentrations is somewhat speeded up.

Due to the cold and dry winter of 1996, and to a lesser extent of 1997, PM levels are elevated (compare leaf #1 in the lower graph of **Figure 20**, and the lower left panel in Figure 21), yielding a lower concentration than was to be expected. An extremely cold winter may lead to PM concentrations, which, on an annual-average basis, are elevated by approximately 4 $\mu g/m^3$. For the *individual* stations this elevation varies from 2 to 7 $\mu g/m^3$.

The fact that PM_{10} concentrations are somewhat lower (~ 2 $\mu g/m^3$) at the end of the nineties (1998, 1999, 2000), can be attributed to meteorological circumstances too. There were fewer days with sub-zero temperatures than usual, and winters were predominantly wet (compare frequencies in the lower left panel of **Figure 21**). At the same time wind speeds in the spring-summer-autumn period were relatively high (compare upper right panel of **Figure 21**).

We conclude that

- the decrease of PM_{10} over the past 10 years accounts for ~11 $\mu g/m^3$. Relative to 1992 this implies a decrease of ~25%;
- this decrease is not influenced by meteorological variability;
- annual concentrations may be *lowered* by ~2 $\mu g/m^3$ for wet and mild years with wind speeds slightly higher than average. Annual concentrations may be *elevated* by ~ 4 $\mu g/m^3$ in years with extreme cold and dry winter months.

5.3 Relation to emissions

We have concluded above that PM_{10} concentrations decreased by ~25% over the period 1992-2001. Now, how can this downward trend be explained? The obvious answer is: by decreasing emissions of PM_{10}. In this paragraph we will couple trends in concentrations and trends in emissions.

Figure 23 shows total anthropogenic emissions of PM_{10} in the Netherlands and surrounding countries (Germany, Belgium, United Kingdom and France). The emissions of primary PM_{10}, SO_2, NO_x and NH_3 are a weighted sum for each country and substance according to their contribution to the annual mean PM_{10} concentration in the Netherlands. The sum of emission-equivalents drops from 318 kTonnes in 1990 to 167 kTonnes in 1999. This is a decrease of almost 50% over the period 1990-1999. Thus, the decrease of emissions is roughly the double of concentrations. How could the difference be explained?

Visser, Buringh and Breugel (2001, Table 34B) have shown that PM_{10} consists of a considerate part of natural dust. This natural contribution was in 1998/1999 roughly 8 $\mu g/m^3$ for the Netherlands as a whole. Contributions are for sea salt (~ 5 $\mu g/m^3$), for wind-blown dust (~ 2 $\mu g/m^3$) and for the Northern Hemisphere background concentration (~ 1 $\mu g/m^3$). If we substract the natural contribution from the annual concentrations in **Figure 23**, the trend decreases from 34.3 $\mu g/m^3$ in 1992 to 23.7 $\mu g/m^3$ in 2001. Relative to 1992 we have a decrease of **31%**.

Thus, also after correction for natural PM sources, emissions still show a stronger downward trend.

From the inferences above we conclude that

- the downward trend in regional PM_{10} concentrations is due to inland and foreign emission reductions;
- the decrease in concentrations is somewhat less pronounced than the decrease in emissions, even if concentrations are corrected for the contribution of natural sources of PM_{10}.

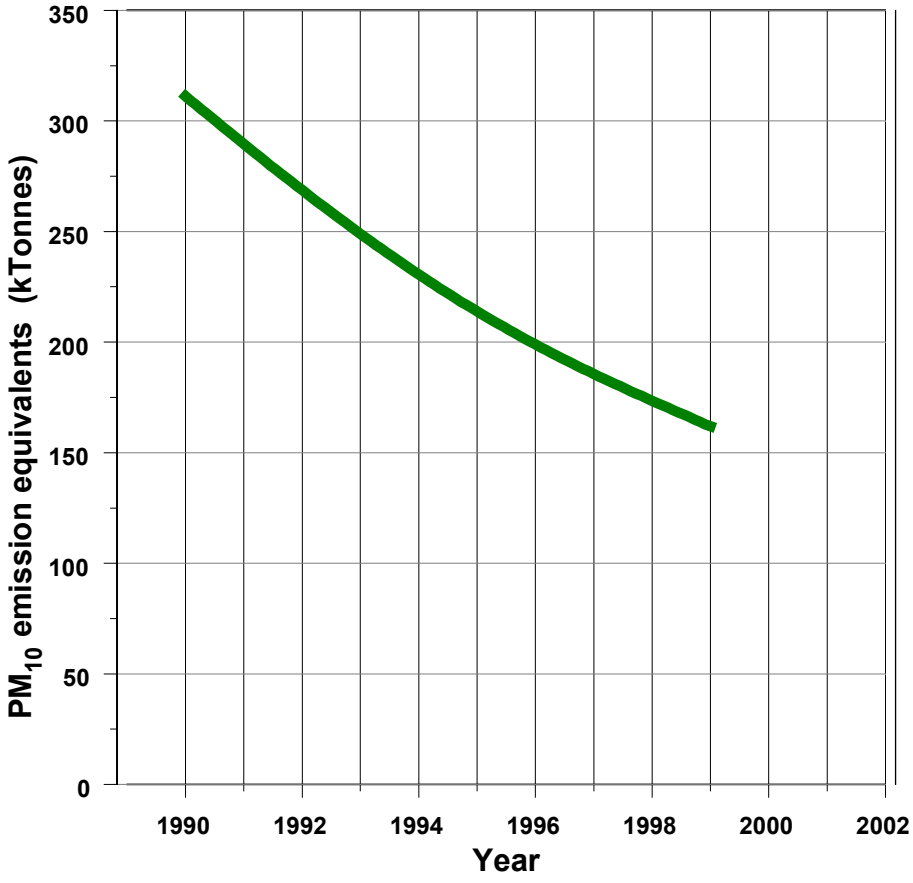

Figure 23 *PM_{10} emission equivalents for the Netherlands, expressed in kTonnes. Emissions were scaled according to anthropogenic contributions of primary PM_{10}, SO_2, NO_x and NH_3 and the surrounding countries Germany, Belgium, United Kingdom and France.*

6. Summary and conclusions

It is well known that a large part of the year-to–year variation in annual distribution of daily concentrations of air pollution is due to fluctuations in the frequency and severity of meteo conditions. This variability makes it difficult to estimate the effectiveness of emission control strategies (Stoeckenius, 1991).

In this report we have demonstrated how a series of binary decision rules, known as Classification and Regresion Trees (CART) can be used to calculate pollution concentrations that are standardized to levels that would be expected to occur under a fixed (reference) set of meteorological conditions. Such meteo-corrected concentration measures can then be used to identify ‘underlying’ air quality trends resulting from changes in emissions that may otherwise be difficult to distinguish due to the interfering effects of unusual weather patterns.

Concentrations of pollutants are influenced by the frequency and severity of meteorological conditions. To remove this influence Regression Tree analysis is an important tool. Adjusted concentrations can now be used more effectively in assigning the influence of control strategies. *Photo: H. Visser*

CART analysis has a number of advantages over other classification methods, such as Multiple Regression or Logistic Regression. First, it is inherently non-parametric. In other words, no assumptions have to made about the underlying distribution of values of the response variable or predictor variables. Thus, CART can handle numerical data that are highly skewed or multi-modal, as well as categorical response variables with either an ordinal or a nominal structure. Second, the relationship between the response variable y_i and predictors **x** may be highly non-linear. Third, the CART results are relatively easy to interpret.

We have refined the Dekkers and Noordijk (1997) methodology for Regression Trees. These refinements comprise:

- checks on the data for outliers, missing values and multi-collinearity among the predictors **x**. The latter tests for instability of the estimated trees;
- checks for transformation of concentrations prior to the estimation of a Regression Tree;
- cross validation of the final tree;
- evaluation of the predictive power of the final tree in relation to rival models. These rival models comprise models based on other statistical principles as well as on a change in sampling time from daily to monthly data.

The Regression Tree approach also has limitations of importance for the interpretation of the Regression Tree results:

- highly correlated predictors induce instability in tree estimates;
- if both response variable y_t and selected predictors $x_{i,t}$ contain long-term trends, the meteo-correction procedure could become sub-optimal. The method cannot uniquely distinguish between emission-induced changes in concentration and meteorologically induced changes.

Although we have given guidelines to detect the occurrence of the situations above, if they occur, results should be presented with care.

We have applied the Regression Tree methodology to nine regional stations of PM_{10} in the Netherlands. Each PM_{10} station consisted of daily data for the period 1992-2001. To couple each station to local meteorological conditions, we divided the Netherlands into five regions and attributed each station to one of these regions. Results are itemized below:

- RT models based on *monthly* concentrations of PM_{10} outperformed those based on *daily* data. Apparently, monthly averaged meteorology is more influenced by large-scale meteorology in Europe, governing periods with extreme concentrations.
- The long-term trend in regional PM_{10} concentrations was not influenced by meteorological variability.
- The concentration trend shows large similarities to trends in *emissions*.
- Anthropogenic *emissions* drop even more rapidly (~50%) than regional *concentrations* corrected for natural emission sources (~30%);
- Due to the cold and dry winter of 1996 (and to a lesser extent that of 1997) concentration levels rose to a great extent. Annual concentrations were elevated by 4 $\mu g/m^3$ (11% of average level in 1996).

References

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., 1984. Classification and regression trees. Wadsworth, Belmont, California.

Buringh, E. and Opperhuizen, A. (editors), 2002a. On health risks of ambient PM in the Netherlands. Executive Summary. Netherlands Aerosol Programme, September 2002.

Buringh, E. and Opperhuizen, A. (editors), 2002b. On health risks of ambient PM in the Netherlands. RIVM report, in press.

Dekkers, A.L.M., 2001. S-PLUS voor het RIVM. Krachtig statistisch software gereedschap. RIVM report 422 516 001.

Dekkers, A.L.M. and Noordijk, H., 1997. Correctie van atmosferische concentraties voor meteorologische omstandigheden. RIVM report 722101 024, RIVM, Bilthoven.

Fritts, H.C., 1976. Tree rings and climate. Academic Press, New York.

Hammingh, P. et al., 2002. Jaaroverzicht luchtkwaliteit 2001. RIVM report, in preparation.

Harvey, A.C., 1989. Forecasting, structural time series models and the Kalman filter. Cambridge University Press, UK.

Jaarsveld, J.A. van, 1995. Modelling the long-term behaviour of pollutants on various scales. PhD Thesis, Utrecht University.

Milieubalans 2002. Het Nederlandse milieu verklaard. Samsom bv., Alphen aan den Rijn.

Milieucompendium 2002. Het milieu in cijfers. Uitgave RIVM/CBS.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, R., Lewandowski, R., Newton, J., Parzen, E. and Winkler R., 1982. The accuracy of extrapolation (time series) methods: results of a forecasting competition. Journal of Forecasting 1, 111-153.

Noordijk, H. and Visser, H., 2002. Correcting air pollution time series for meteorological variability. Refinements to the method of Stoeckenius. Atm. Environment, in preparation.

Ripley, B.D., 1996. Pattern recognition and neural networks. Cambridge University Press.

Saltelli, A, Chan, K. and Scott, E.M., 2000. Sensitivity analysis. Wiley & Sons Ltd., Chichester, UK.

S-PLUS 2000. Modern statistics and advanced graphics. Guide to statistics, vol I (Chapter 12). MathSoft Inc., Seattle WA.

Stoeckenius, T., 1991. Adjustment for ozone trends for meteorological variations. In Tropospheric Ozone and the Environment 3. Papers from an international speciality conference, Berglund, R.L., Lawson, D.R. and Mckee, D.J. (ed.), Air & Waste Management Association, Pittsburgh PA.

Venables, W.N. and Ripley, B.D., 1997. Modern applied statistics with S-PLUS. Springer-Verlag, New York Inc.

Visser, H., 2002. Correcting air pollution time series for meteorological variability. Regression Tree software implemention in ART. RIVM Memorandum, in preparation.

Visser, H., Buringh, E. and Breugel, P.B. van, 2001. Composition and Origin of airborne Particulate Matter in the Netherlands. RIVM report 650010 029, RIVM, Bilthoven.

Appendix A S-PLUS script RTAonPseudoseries

```
#############################################################################
#
# S-PLUS-script for perfoming a full Regression Tree analysis as an 8-step
   procedure.
# Original data are copied to dataframe 'aa.dat' for convenience.
#
# Programmer: H.Visser (CIM-RIVM)
# Date: 11 June, 2002
#

#############################################################################
#
# Step 1.  Descriptive analysis of data.
#
aa.dat <- Pseudo5[-1:-365,]

m <- menuDescribe(data = aa.dat, variables = "<ALL>", grouping.variables =
    "(None)",
max.numeric.levels = 10, nbins = 6, min.p = T, first.quant.p = F, mean.p =
    T, median.p = F,
third.quant.p = F, max.p = T, nobs.p = T, valid.n.p = T, var.p = F, stdev.p
    = F, sum.p = F,
factors.too.p = T, print.p = T, se.mean.p = F, conf.lim.mean.p = F,
    conf.level.mean = 0.95,
skewness.p = F, kurtosis.p = F)
f.print(m)
#
guiPlot( PlotType = "Scatter Matrix", DataSet = "aa.dat", Columns = "Index,
    Tgem, Tmin, Tmax,
     RHgem, Regenmm, Regenduur, Pgem")
graphsheet()
guiPlot( PlotType = "Scatter Matrix", DataSet = "aa.dat", Columns = "Index,
    Stralgem, PercStral,
     Windr, Windkr, Pasqudag, Pasqunacht")
#
luvo <-
    c("Tgem","Tmin","Tmax","RHgem","Regenmm","Regenduur","Pgem","Stralgem",
  "PercStral","Windr","Windkr","Pasqudag","Pasqunacht")
#
```

```
#########################################################################
#
# Step 2. Here we make 2 plots from which we can judge if a transformation
#         on the original concentrations is needed. If points in the Range-
    mean plot
#         show an increasing tendency, this points to a log-trafo or
#         to a division by an estimated polynomial trend.
#         N.B.: in the first line the dataframe 'aa.dat' with y en x values
#               is copied to 'x.dat'. We only use the column with variable
#               JJJJ and the y-variable. First 'JJJJ', then 'yvar'.
#
x.dat <- aa.dat[,c(4,1)]
len <- length(x.dat[,1])
x.dat <- cbind(1:len,x.dat)
names(x.dat)[1] <- "time"
x.dat[1:10,]
Polyorder <- 2
x.lm <- lm(x.dat[,3] ~ poly(x.dat[,1], Polyorder), na.action = na.exclude)
trend <- predict(x.lm)
plot(x.dat[,1], x.dat[,3], xlab= "Time", ylab= "Concentration")
lines(x.dat[,1], trend, col = 4, lwd = 3)
title(paste("Original data and ", Polyorder, "order trend"))
#
x.mean <- aggregate.data.frame(x.dat[,3],x.dat[,2],mean,na.rm=T)
x.mean
x.var <- aggregate.data.frame(x.dat[,3],x.dat[,2],var,na.method="omit")
x.var
x.sd <- sqrt(x.var$x)
x.ave <- x.mean$x
plot(x.ave,x.sd, xlab= "Annual mean concentration", ylab= "SD of annual
    concentration")
title("Range-mean plot for original concentrations")
#

#########################################################################
#
# Step 3.   First analysis of the Regression Tree and subsequent pruning of
    this tree.
#
f.RTAnew(aa.dat, "Pseudo5", "Index",luvo,1,80,40,0,0)
#

#########################################################################
#
# Step 4.  Calculation of diagnostics and meteo-corrected concentrations
   for the optimal tree.
#
#
f.RTAfreq(aa.dat,"Pseudo5", "Index",luvo,1,80,40,8,2001,0,0,0)
#

#########################################################################
#
# Step 5.  Cross validation of the optimal tree. The years that are omitted
    in the estimation of the tree,
#          are in 'cvyears'.
#
cvyears <- c(1995,1999)
f.RTAcrossval(aa.dat,"Pseudo5", "Index",luvo,1,80,40,8,cvyears)
```

```
#
################################################################################
#
# Step 6.  Three trellis graphs are generated for diagnostic checking of
    the estimated tree.
#
#
b.dat <- na.omit(Pseudo5.Index.met1.dat)
graphsheet()
barchart(nleaf~n|paste("year=",as.character(JJJJ)),b.dat)
title("Number of days n, as a function of leafnr, split up for each year")
barchart(JJJJ~n|paste("leafnr=",as.character(nleaf)) ,b.dat)
title("Number of days n, as a function of yearnr, split up for each leaf")
histogram(~residual|paste("leafnr=",as.character(nleaf)), b.dat)
title("Distribution of RT residuals for all years, split up for each leaf")
#

################################################################################
#
# Step 7.  A time-series plot of the RT predictions and the meteo-corrected
    concentrations can be generated by starting script 'RTAplotPredictions'.
#

################################################################################
#
# Step 8.  Finally we perform a Multiple Regression analysis on the same
    data as we use for RTA: 'aa.dat'.
#
f.MultRegr(aa.dat,"Pseudo5","Index",luvo,1,2,0,0)    #
#


################################################################################
#
# The end.
#
################################################################################
```