Developments in Statistics and Methodology
A. Ferligoj and A. Kramberger (Editors)
Metodološki zvezki, 9, Ljubljana: FDV 1993

An Optimization Approach to the Problem of Short Reference Year

Katarina Košmelj[1], Damjana Virant[2], Matjaž Jeran[3]

Abstract

For a chosen location, the short reference year (SRY) is a time series of representative days for given time periods. Very often the calendar year is divided into 36 decades and SRY is a time series of hourly values for 36 representative days, each representing a particular decade. SRY is used as input data for different simulation models (e.g. agrometeorological models, energy consumption models etc).

This paper presents a simple and efficient algorithm to the problem of SRY. The methodology is based on a two-step optimization approach. In the first step for each time period the representative day for each meteorological variable under study is found. In the second step the representative day for the whole set of variables is searched for. The criterion functions used are based on statistical parameters.

The methodology was applied on a set of two meteorological variables: temperature and sunshine duration. The data for the Ljubljana station were supplied for the period 1960-1979 as simultaneous hourly values. Results for two decades of the SRY are presented.

1 Introduction

In the past, weather conditions for a particular site were usually described by mean daily (monthly, yearly) values of several meteorological variables. Nowadays, very often hourly values are required, in particular for the computer simulations of processes in the environment (plant growing, solar energy consumption etc). Hourly values for a set of meteorological variables require a large computer storage. This problem can be avoided by reducing or by condensing the data. The Test Reference Year (TRY) and its special case the Short Reference Year (SRY) are two ways of doing that.

For each variable under study the TRY is a time series of "representative" days. The length of the TRY is 365 days, each day being described by hourly values.

[1]Biotechnical Faculty, University of Ljubljana, Jamnikarjeva 101, 61000 Ljubljana, Slovenia
[2]Hydrometeorological Institute of Slovenia, Vojkova 1b, 61000 Ljubljana, Slovenia
[3]ORACLE SOFTWARE, Dunajska 160, 61000 Ljubljana, Slovenia

Several authors have dealt with this problem (Cehak 1975, Lund 1976, Hoogen 1976, Boehe et al. 1979, Kletter 1982, Kajfež-Bogataj and Hočevar 1985, Kajfež-Bogataj and Hočevar 1985a, Zupančič and Pristov 1987, Babilon 1989, Kajfež-Bogataj 1990).

For the SRY the calender year is divided into periods. The length of a period depends on the nature of the problem. Usually it is one month, 10 days (decade) or one week which means 12, 36 or 52 representative days described by hourly values. Each day is representative for a particular period.

In our setting two different approaches were used to construct the TRY. The first (Kajfež-Bogataj 1986, Hočevar and Kajfež-Bogataj 1986) was based on the mean monthly values of 23 meteorological variables. Nonparametric statistical tests were used to determine the representative months. This approach was applied for Ljubljana only. The second approach (Zupančič and Pristov 1987) was based on the mean monthly values of 26 meteorological variables, numerical taxonomy was used to determine the representative month on the base of Euclidean distances. The TRY was applied on the data from Ljubljana, Maribor and Koper.

In this paper we present a simple approach to construct the SRY. It is based on a two-step optimization technique. In the first step the representative days for each variable separately are determined. In the second step the representative days for the whole set of meteorological variables is searched for. In each step a criterion function is prescribed in advance and its optimum calculated. In illustration we present the SRY for a set of two meteorological variables: air temperature and sunshine duration. The data were measured in the period 1960-79 in Ljubljana.

2 Methodology

2.1 Objective

The objective of our work was to develop a simple and efficient algorithm to construct the SRY. To obtain a sensible solution of the SRY problem the following conditions are imposed:

- the variables under study are quantitative, they may be continuous or discrete, having different probability distributions;
- some data may be missing;
- the approach should be flexible for different period lengths;
- the representative days should be independent of each other;
- the results should be acceptable from the physical point of view.

2.2 Input data

The input data $\underline{X}$ can be presented as M three-dimensional data matrices

$$\underline{X} = (X^1_{h,d,y}, X^2_{h,d,y}, ...X^M_{h,d,y})$$

$\underline{X}$ denotes the set of M input variables $X^1, X^2 ... X^M$. Subscripts h, d and y denote: hour $h = 1...24$, day $d = 1...D$ and year $y = 1...Y$. For example, if data for

30 years present the input, then $D = 365$ or 366 (for leap-years) and $Y = 30$, for each variable under study.

2.3 Output data

From input matrices $\underline{X}$ we want to obtain M two-dimensional output matrices $\underline{Y}$:

$$X^i_{h,d,y} \longrightarrow Y^i_{h,p}, i = 1...M$$

Subscripts $h = 1...24$ and $p = 1...P$ denote hours and periods. Each period is L days long. If the SRY for decades ($L = 10$) is constructed on the input data mentioned above, the data are condensed to M output matrices with $P = 36$. In that case the input data are reduced for the factor $10 * Y = 300$ approximately .

2.4 Optimization procedure

2.4.1 Univariate analysis

First of all, for each variable X^i the meaning of the representative day for a period must be defined. We impose the following definition: the representative day for the chosen period is that day **in the input data** which is the most similar to the "statistical" day for this period.

Statistical day for the period: The statistical day $X^i_p = X^i_{h,p}, h = 1...24$ is defined to summarize the input data for a particular period p over all the years. To define the statistical day a measure of central location can be used, as for example mean, median, mode. Its choice depends on the empirical probability distribution of the variable under study.

For example, the "mean day" for X^i for the period p is a series of mean hourly values $\overline{X^i_{h,p}}$

$$\overline{X^i_{h,p}} = \frac{1}{LY} \sum_{d \in p} \sum_{y=1}^{y=Y} X^i_{h,d,y}, h = 1...24$$

Representative day for the period: To select the representative day a criterion function must be defined. For each day in the input data the criterion function measures the dissimilarity between that day and the statistical day. There are several possibilities of doing that. Measures of dissimilarity from cluster analysis, in particular the measure of dissimilarity for time series, are appropriate (Košmelj and Batagelj 1990). In illustration we present the criterion function D in the form of the modified Euclidean distance.

$$D(X^i_{d,y}) = \sqrt{\sum_{h=1}^{24} q_h (X^i_{h,d,y} - X^i_{h,p})^2}, d \in p, y = 1...Y$$

$q_h, h = 1...24$ are the weights which express the relative importance of a particular hour during the day. Usually $q_h = 1$ for $h = 1...24$.

The output is the day $Y_p^i = Y_{h,p}^i, h = 1...24$ with the minimal value of the criterion function

$$D(Y_p^i) = min_{d \in p, y=1...Y} D(X_{d,y}^i)$$

To summarize the univariate analysis: P statistical days are determined first. Then for each period L*Y values of the criterion function are calculated and the minimum is searched for. P optimizations are carried out for each variable.

That procedure is repeated M times, for each variable separately.

2.5 Multivariate analysis

In the univariate analysis for each variable under study and for each period under study a representative day is obtained. For one period the representative days for several variables do not necessarily coincide.

From the physical point of view it would be reasonable to find out **one** representative day for the whole set of variables, for each period under study (global representative day). Therefore the new criterion function F is to be defined which takes into account all M variables simultaneously.

The multivariate criterion function F is expressed as a linear combination of the normalized criterion functions $D^*(X_{d,y}^i), i = 1...M$.

$$D^*(X_{d,y}^i) = \frac{D(X_{d,y}^i)}{max_{d \in p, y=1...Y} D(X_{d,y}^i)}$$

$$F(X_{d,y}^i) = \sum_{i=1}^{M} w_i D^*(X_{d,y}^i)$$

The weights w_i are defined according to the relative importance of the variable X^i in the set of variables.

For each period under study the global representative day is the one with the minimal value of the criterion function F. Finally P global representative days are obtained.

To perform stable sorting bubble sort method was used (Kozak 1986).

3 Illustration

The methodology was applied on the data set of only two meteorological variables: air temperature and sunshine duration in Ljubljana for the period 1960-79. These two variables describe quite well the general thermal weather conditions, which are often required when modelling solar energy consumption, indoor climate controlling etc. For illustration, results for two decades (second decade in June and in December) will be presented.

3.1 Description of the data

Air temperature was measured at the hours ($1^h, 2^h, ...24^h$) and is expressed in Centigrades as instantaneous value. The Campbell-Stokes sunshine recorders were used to measure sunshine duration. Sunshine duration is an interval variable with the values from 0 to 1.0. Its value represents the proportion of sunshine duration during a particular hour. The data were obtained from the Hydrometeorological Office of Republic Slovenia. No data were missing.

3.2 Presentation of the results

Both variables are continuous, but the shape of their empirical probability distribution is different. Air temperature has the empirical probability distribution which is symmetric (Figure 1, 2). Therefore for the statistical day the mean day was chosen. On the other hand, the empirical probability distribution for sunshine duration is in U shape or J shape (Figure 3, 4). Therefore the median day was calculated for each period.

Euclidean distance with the weights $q_h = 1$ was used.

For air temperature the statistical days (mean day, median day) and the representative day lie on a quasi-sinus curve (Fig. 5). In both cases the median days and mean days closely coincide. The representative day is more similar to the statistical days in June than in December.

The statistical days and the representative day for sunshine duration in the second decade of June and December are plotted on Figure 6. There exists a big difference between the median day and the mean day. Due to the showed distribution of sunshine duration the mean day is not similar to any historical day. The differences between the statistical and the representative days for sunshine duration are greater than for air temperature.

It was interesting to find out that for the second decade in December the representative day for sunshine duration was not uniquely determined. The hourly values for the median day were 0 constantly (Figure 6), so several days with no sunshine could be taken as the representative day for this period.

The representative days for the second decade of June for air temperature and for sunshine duration did not coincide.

In the calculation of the criterion function F both variables were regarded, for the sake of simplicity, as equally important.

For some decades of the SRY the global representative day was identical to the sunshine duration representative day, for some decades to the air temperature representative day (winter periods) and sometimes it was different from both. In the latter case, it may be interesting to mention that the global representative day for those periods was always one of the first ten days, with the minimal values of criterion function, for each variable.

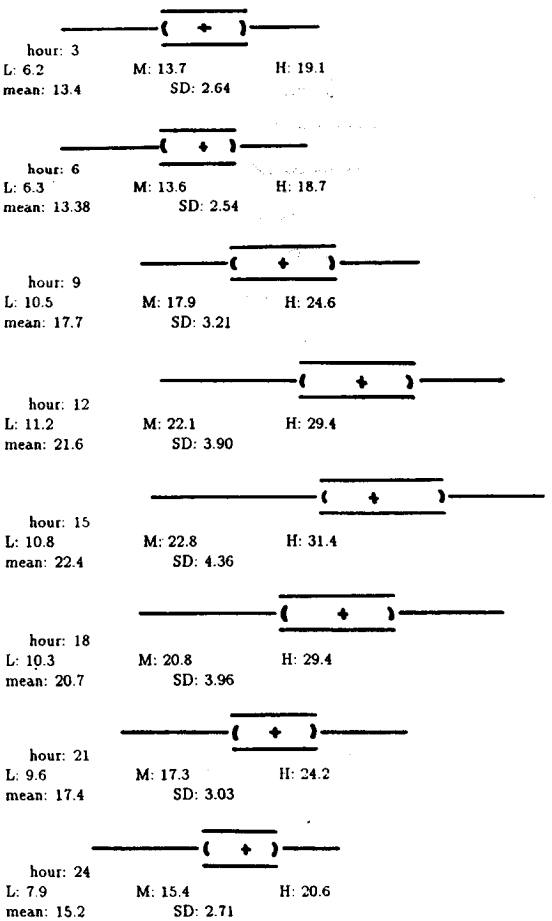

Figure 1: Box and whisker plot and descriptive statistics for the hourly values $(3^h, 6^h, ..., 24^h)$ of air temperature for the second decade of June, L - minimum, M - median (+), H - maximum, SD - standard deviation, * - outlier, (- first quartile,) - third quartile.

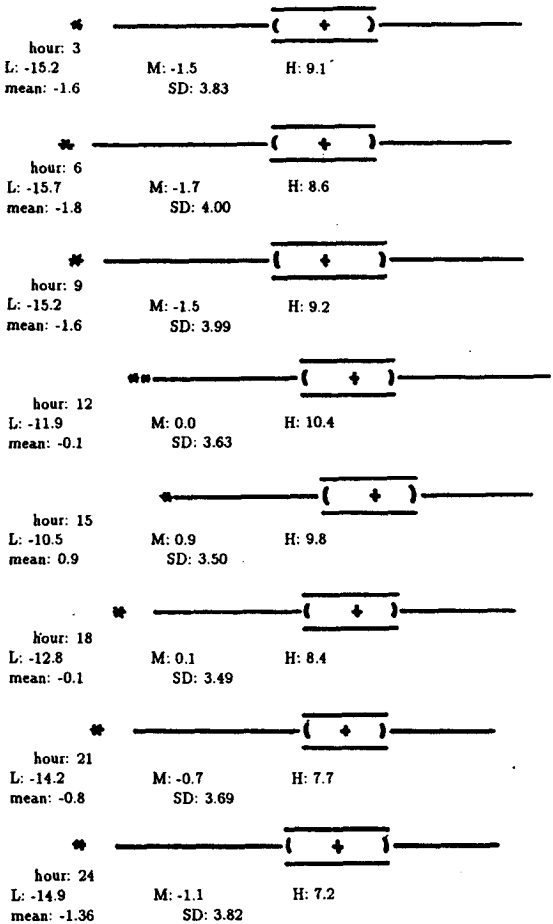

Figure 2: Box and whisker plot and descriptive statistics for the hourly values ($3^h, 6^h, ..., 24^h$) of air temperature for the second decade of December, L - minimum, M - median (+), H - maximum, SD - standard deviation, * - outlier, (- first quartile,) - third quartile.

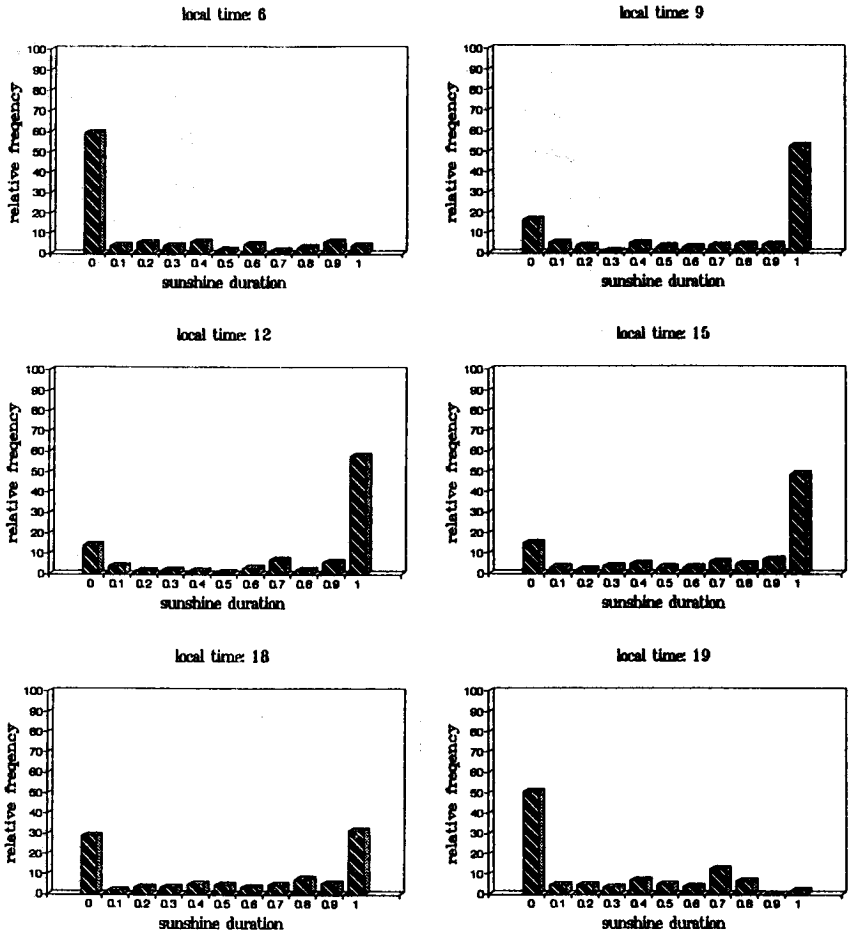

Figure 3: Histograms for the hourly values ($6^h, 9^h, ...19^h$) of sunshine duration for the second decade of June.

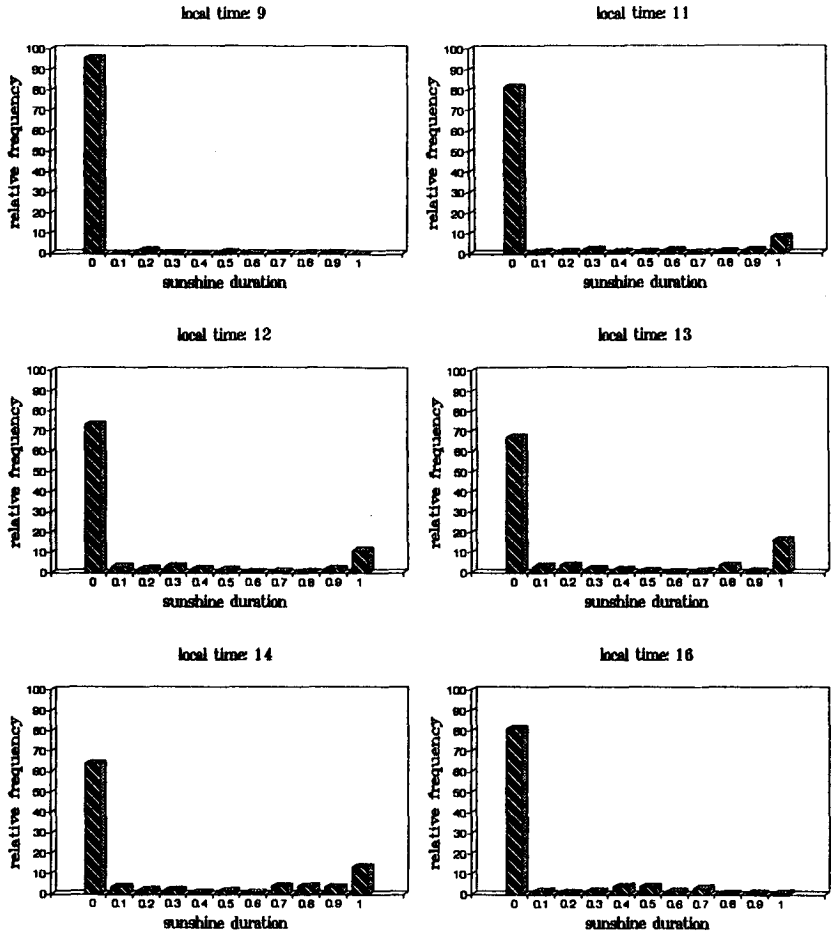

Figure 4: Histograms for the hourly values $(9^h, 11^h, ..., 16^h)$ of sunshine duration for the second decade of December.

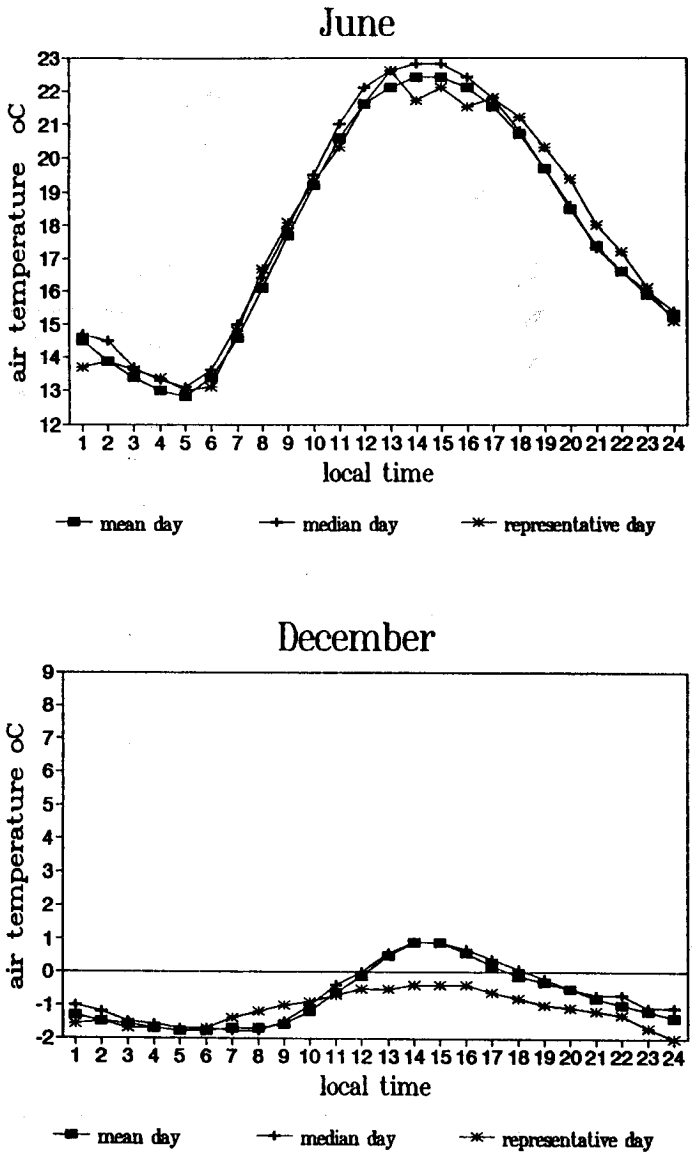

Figure 5: Hourly values of air temperature for the statistical days (mean day, median day) and for the representative day for the second decade of June and for the second decade of December.

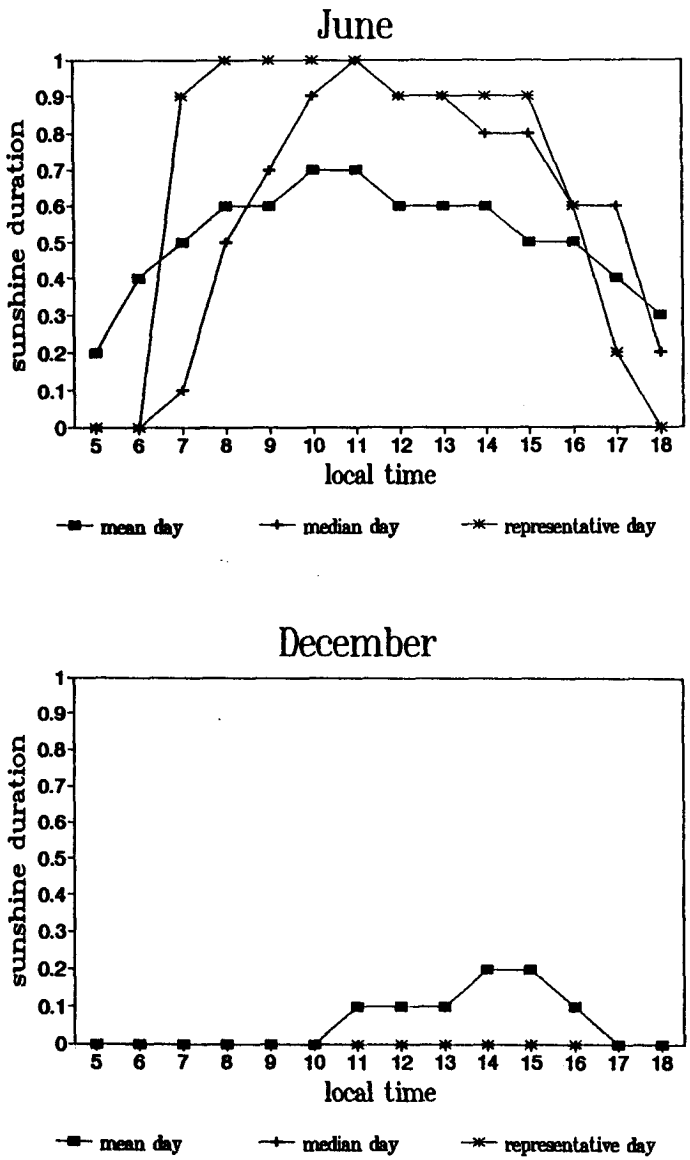

Figure 6: Hourly values of sunshine duration for the statistical days (mean day, median day) and the representative day for the second decade of June and for the second decade of December.

4 Conclusions

The presented approach is very simple and flexible and it accounts for all the requirements stated in the methodology section.

We can choose any number of meteorological variables with different empirical probability distributions of hourly values. For each variable a criterion function can be formulated independently. The representative days for a particular variable are independent of each other. The length of the period is arbitrary. The weights should in principle be according to the needs of the user of the SRY. For example: if the SRY is used for illumination studies, the sunshine duration seems to be more important, if the SRY is needed for studies of heat flow through the northern walls of houses, air temperature is more important. The weights should be determined independently from the described statistical procedure. The problem of missing values can be avoided by adjusting the values of the weights. The representative days are actual historical days and thus the results are "physically acceptable".

As a side result the extreme days i.e. the days with the maximal value of the criterion function D are obtained. This information can be very useful in those fields of science where the extreme conditions are of interest (for example building etc.).

Interdisciplinary approach in this kind of scientific activity is of great importance. The researches must have good knowledge of both the physical and the statistical characteristics of variables.

Only quantitative variables can be used. However, when the SRY is required this is not a disadvantage, because variables are rarely qualitative.

References

[1] Babilon, V. (1989): *Standardna meteorološka godina za Novi Sad.* Magistarski rad, Univerzitet u Bratislavi.

[2] Boehe, A. W., von Paasen, A. H. C., and Jong, O. (1979): *Ein synthetisches Referenzjahr fur Energiebedafsberechnungen.* Haustechnik, Bauphysik, Umwelttechnik, 100 p.

[3] CEC, Commission of the European Communities (1986): *Short Reference Years for CEC Countries.* EUR 10663, 31 p.

[4] Cehak, K. (1975): TRY - Ein Test-Reference-Jahr fur Energieberechnungen in der Raumphysik. *Wetter und Leben*, 27, 3/4, 254-258.

[5] Hočevar, A., and Kajfež-Bogataj, L. (1986): *Urne vrednosti standardnega meteorološkega leta za Ljubljano.* BF - Agronomija, Poročilo za RSS, 83 p.

[6] Hoogen, van de H. (1976): Ein Referentiejaar vor Nederland. *Klimaatbeheersing*, 5, 10.

[7] Kajfež-Bogataj, L., and Hočevar, A. (1985): Osnove za oblikovanje standardnega meteorološkega leta. In: *Zbornik kongresa Rave*, Portorož, 179-197.

[8] Kajfež-Bogataj, L., and Hočevar, A. (1985a): Oblikovanje in uporaba standardnega meteorološkega leta. *Zbornik BF*, **45**, 9-21.

[9] Kajfež-Bogataj, L. (1986): Non parametric method for construction of test reference year. In: *Proceedings of Third Intern. Conf. on Statistical Clymatology*, Vienna, 430-436.

[10] Kajfež-Bogataj, L. (1990): Testno referenčno leto in kratko referenčno leto - izbor meteoroloških podatkov pri modeliranju vpliva okolja za razne namene. In: *Zbornik referata med. kong. Energija i zaštita čovjekove okoline*, Opatija, 247-254.

[11] Kletter, B. (1982): Eine osterreichische Version eines Test- Referenz-Jahres fur die Anwendung in der Bauphysik. *Arch. Met. Geoph. Biokl.*, B, **30**, 3, 283-301.

[12] Košmelj, K., and Batagelj, V. (1990): Cross-sectional Approach for Clustering Time Varying Data. *Journal of Classification*, **7**, 99-109.

[13] Kozak, J. (1986): *Podatkovne strukture in algoritmi.* Ljubljana: DMFA.

[14] Lund, H. (1976): Test Reference Year Weather Data for Environmental Engineering and Energy Consumption in Buildings. Paper to *CIB-17 Meeting*, London.

[15] Lund, H. (1985): *Short Reference Years and Test Reference Years for EEC Countries.* Technical University of Denmark.

[16] Virant, D., and Hočevar, A. (1990): Primerjava karakterističnih meteoroloških parametrov različno oblikovanih standardnih meteoroloških let za Ljubljano. *Zbornik referata med. kong. Energija i zaštita čovjekove okoline*, Opatija, 431-443.

[17] Zupančič, B., and Pristov, J. (1987): Metoda in izbor referenčnega meteorološkega leta, *Razprave-Papers*, **29**, 1, Ljubljana: DMS, 3-12.