THE AUTOMATIC DIGITIZATION OF TIME SERIES RECORDED ON GRAPH PAPER SUPPORTS

Robertomassimo LEONARDI [1], Tito MONTEFINALE [1], Vincenzo MALVESTUTO [1], Olivia TESTA [1], Maria Carmen BELTRANO [2]

[1] Istituto di Scienze dell'Atmosfera e del Clima, CNR, Via Fosso del Cavaliere 100, Rome, Italy
tel. +39 06 49934264, r.leonardi@isac.cnr.it, v.malvestuto@isac.cnr.it

[2] Ufficio Centrale di Ecologia Agraria, CRA, Via del Caravita 7/A, Rome, Italy
tel. +39 06 69531205, beltrano@ucea.it

ABSTRACT

The acquisition of meteorological data, such as rainfall, temperature, humidity, pressure, wind speed and direction, done using mechanical recorders, used in the past to occur by means of a pen-nib, linked to an instrument transducer, drawing a continuous track on graph paper. When one needs to transfer the information contained in these tracks into files, manual techniques or hand-digitiser techniques can be used, but such methods tend to adds random relative errors valued at 5-7%, depending on the training and the tiredness of the operator.

In this paper a completely automatic method to accomplish this task is presented and applied to hundreds of tracks taken from historical databases made available by the Italian Meteorological Service.

The method has been implemented in a software, called *DigiGraph*, written in C – language and running under Extended DOS, consisting of a collection of routines able to automatically read and digitize scanner images of pluviograms, thermograms, udograms or barograms. The use of this software makes thus possible to recover within reasonable times the vast information stored in the voluminous paper archives existing in many organisations and institutions, that have been accumulated throughout many years and in some cases for more than a century, whose management and conservation is itself a serious issue.

INTRODUCTION

A meteorological station equipped with electronic sensors and automatic recording on computer files represents the modern system of data acquisition. This makes obsolete the traditional mechanical recorders based on pen-nibs leaving continuous tracks on purposely designed graduated papers, with the further advantages that sampling times as small as few minutes are easily obtainable and the data are immediately available for further processing and analysis.

However, since the traditional data still wait, typically stored in large quantities inside huge dusty archives, to be patiently digitized, and since in many countries the observational networks engaged in environmental monitoring continue to acquire meteorological data with the above mentioned traditional methods, there still are lots of data uneasy to handle for the purposes of the meteorological research.

The factors limiting the use of these data acquired with traditional methods are essentially two:

1. There are several possible errors in the manual procedure: errors of reading, of deciphering the track, of exactly copying the read values etc., and anyway the operations are inaccurate, lengthy and tedious. The exact reading of the track is difficult because the time scales are not easily readable since the underlying grid is curvilinear. Furthermore, for some quantities, like relative humidity, the vertical scales are nonlinear. Therefore, there is a high degree of subjectiveness in the manual digitization of the data, the results depending on the experience and on the training of the operator and often even on his personal insight.
2. The conservation of the cartograms represents a second critical factor. In fact, as years pass by, the paper gets dusty, spoiled and worm-eaten, while the tracks fade and become less and less readable, so that the longer is the delay in digitizing the cartograms, the more difficult is the retrieval of the recorded data and of the meteorological information stored therein.

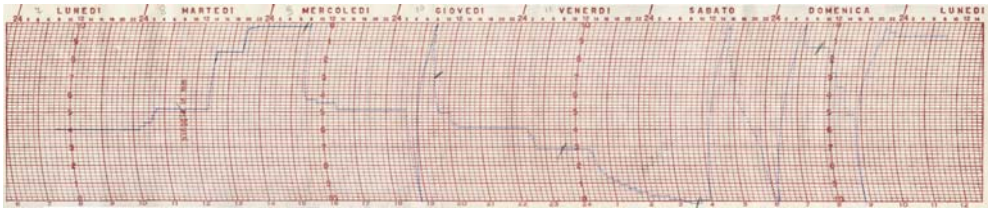
As a matter of fact, several software dedicated to data digitization are available which allow the transcription of the paper-recorded data onto text files, after the acquisition of the tracks by a scanner as an image file. But these commercial programs are typically only semi-automatic and require a fundamental engagement of the operator (frequent mouse-clicking along the track) in order to

reconstruct the time evolution of the quantity under examination. These techniques require rather long acquisition times and do not warrant a complete and accurate data transcription.

THE "DIGIGRAPH" SOFTWARE

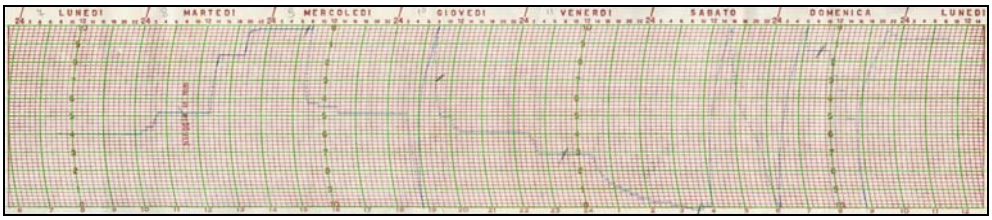
One of us (R.L.) has conceived, prepared and tested a software, called *DigiGraph*, dedicated to the completely automatic reading of scanned images of records of precipitation, temperature, relative humidity and pressure. The same software also provides the immediate storage of the resulting data on text files according to a previously user-defined format. The method has been implemented in a code (consisting of many thousands of statements) written in C-language and running under Extended DOS. The software *DigiGraph* actually is a large collection of routines whose combined action yields the automatic conversion of scanner-digitised pluviograms, thermograms, udograms or barograms into files of simple numeric format. The whole procedure can be divided in few steps:

- A cartogram is converted into a file by means of a scanner as a true-colour (24 bit) image with a resolution of 150 to 200 depending on the size and quality of the cartogram.



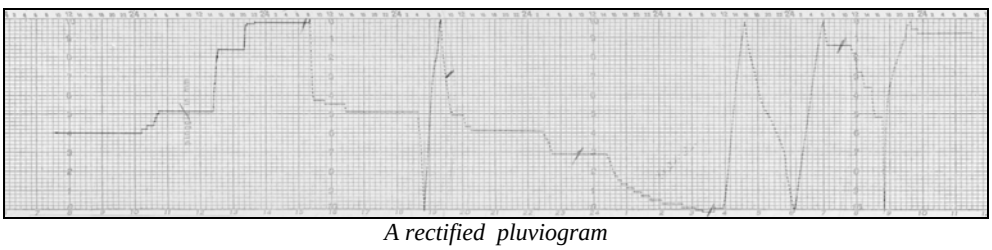
Original pluviogram - Allumiere (RM) (07/02/1983 - 14/02/1983)

- A configuration file is associated to each image file, containing: name of the station, time span of the recording and, optionally, date and time of the beginning and of the end of the track, and, if required, other information useful to check the global quality of the data rendering (for instance, total water volume collected by a pluviometer in the whole week, or minimum and maximum temperature within the given thermogram, etc.)
- Then the program is run, and the user is asked to select first the desired image via a menu. When this is done, the corresponding image is displayed.
- The user checks that the image possesses the proper requisites (no multiple entangling tracks, no too long gaps in the track, visibility of the track, and so on) to start the elaboration. The user at this point can choose either to proceed or to discard the cartogram.
- The program detects and vectioralizes the curvilinear grid with a suitable algorithm based on a 2-D Fourier transform.
- The vectorialized grid is presented superimposed to the image, and the user is left with the final decision of rejecting the grid retrieval or accepting it and proceeding further.



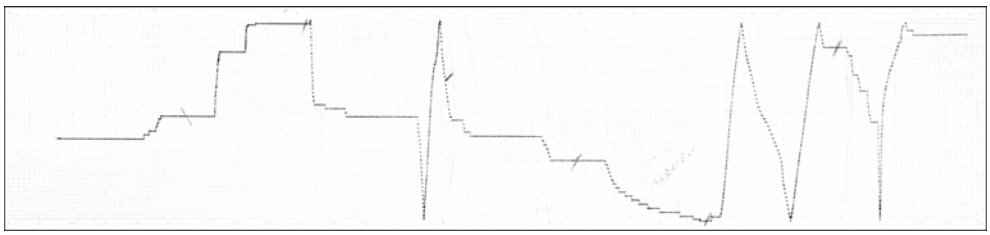
Check of the curvilinear grid (green superposed line) detected by the software.

- A transformation from cylindrical to Cartesian coordinates is applied to the whole image, producing the simultaneous rectification of the grid and the track.



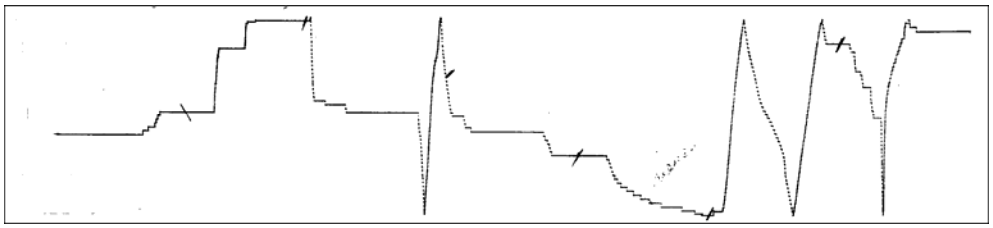
A rectified pluviogram

- A specific algorithm, based on colorimetric and morphological analyses, has been developed in order to identify and remove spots and artifacts marring the cartogram and to optimize the contrast between the tracks and their background.



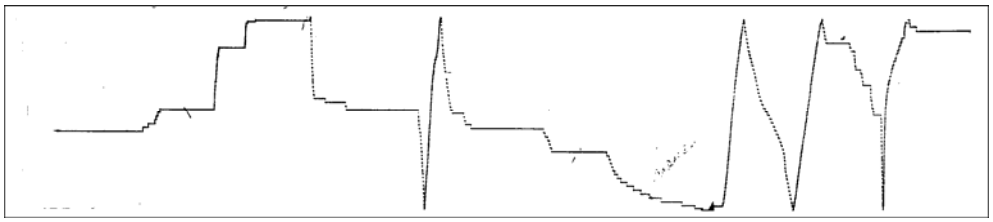
Contrasted image evidencing the track while attenuating the background with the grid

- A double-threshold criterion, based on a variant of the maximum entropy algorithm (modified Kapur method), is applied to the optimised image forcing to black all the dark parts of the image (namely, the track and some spurious artifacts) and to white the light parts (the background and the grid).



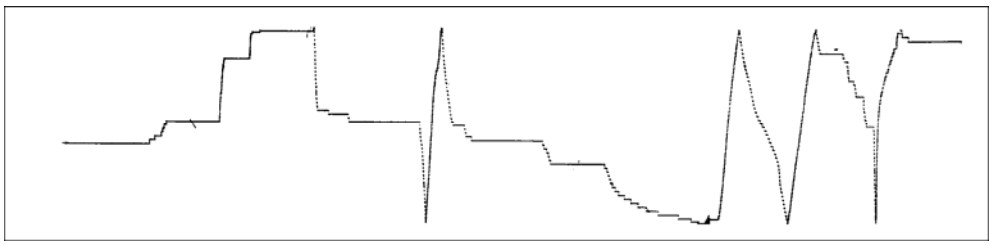
Reduction of the contrasted image to 2 levels only (black and white)

- At this stage the image consists only of the track, with its possible gaps, overlapping near sharp peaks, bifurcations (due to embedded spurious objects like pencil marks), and of other artifacts like ink blots, spots of various nature, often hand-made notes or stamps, added by mindless operators, often in the same colour as the track. At this point, a whole collection of algorithms, by working jointly, make it possible to recognize and to clear all the spurious objects leaving the track finally ready for the actual digitization.



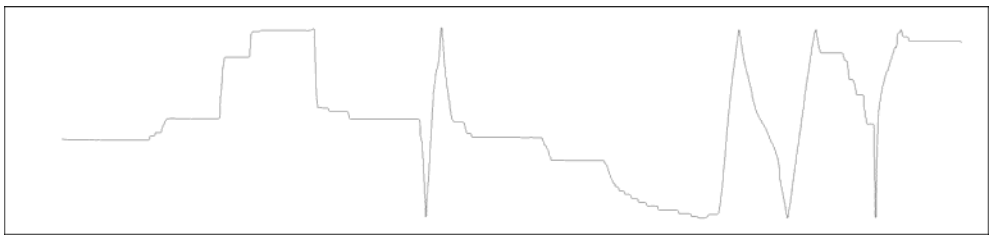
Removal of spurious artifacts likes spots and pencil marks

- After this stage the track may still contain gaps that must be filled in, unless they are too wide (in the latter case they are left as they are). The process of reconstructing the short missing parts is done by an iterative algorithm of growth/thinning/refining of the image which produces a vector containing all and only the points lying on the axis of the track.



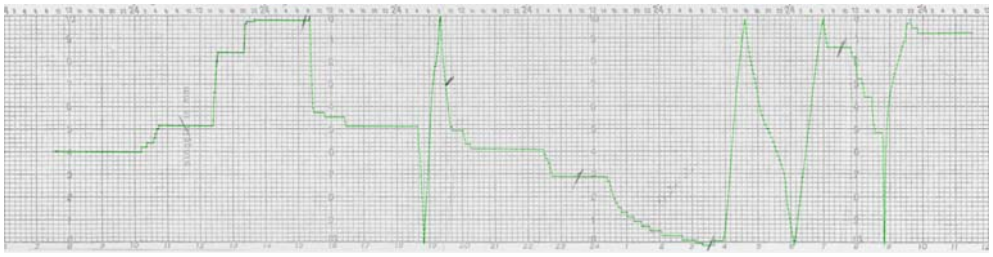
First cycle of reconstruction of the track

- At last one gets the reconstructed track

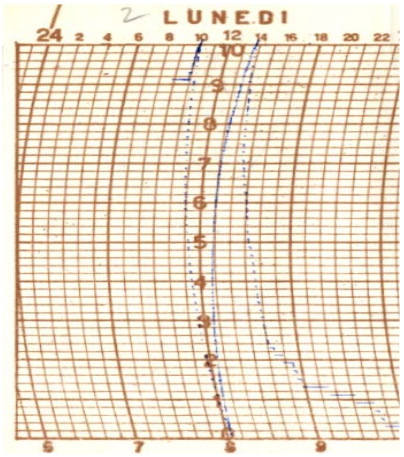


Reconstructed track

- The vectorialized image of the track is displayed superimposed onto the original image. If the result is accepted by the user, the procedure ends with the generation of a data file containing the cumulated precipitation recorded every, say, 5 minutes as a function of time.



The track detected by the software is superimposed in green on the original rectified image



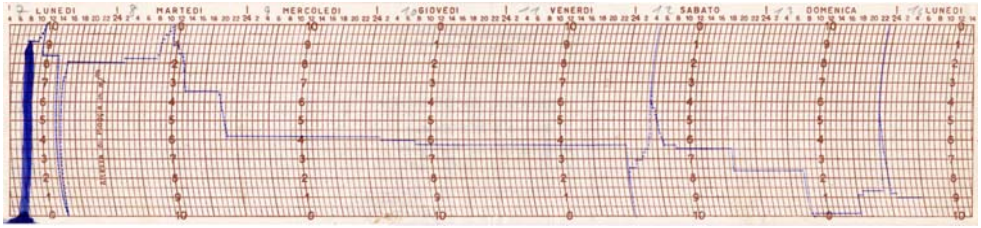
Station: Allumiere (VT)
Starting Date 02/03/1981
Total water collected: 4.4 lt

Data	ora	Prec.tot (mm)
02/03/1981	09:00	0.0
02/03/1981	09:05	0.0
02/03/1981	09:10	0.0
02/03/1981	09:15	0.0
02/03/1981	09:20	0.0
02/03/1981	09:25	0.0
02/03/1981	09:30	0.0
02/03/1981	09:35	0.0
02/03/1981	09:40	0.0
02/03/1981	09:45	0.0
02/03/1981	09:50	0.0
02/03/1981	09:55	0.8

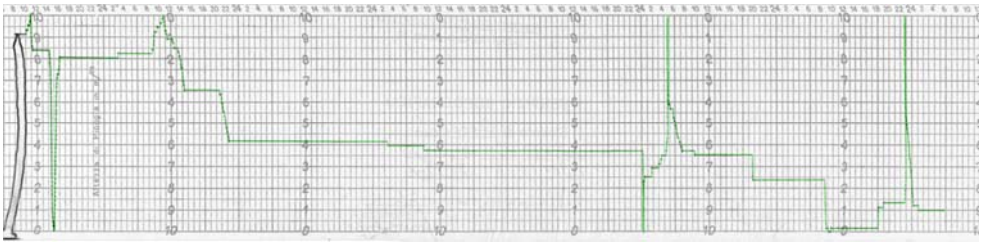
*The initial segment of data in the output data file generated by **DigiGraph***

The procedure described above has proved to be able to cope successfully with more than 95% of a multivaried sample of about 360 cartograms, of which 200 pluviograms coming from the Allumiere (VT) station, and other 52 thermograms, 52 udograms and 52 pluviograms coming from the Collegio Romano station. Experience shows that the conditions to be satisfied for a successful processing are:

- The conservation status of the paper support must at least allow the reading by eye of the track
- The track should not present gaps exceeding two hours (in a weekly cartogram)
- There should not be multiple tracks entangled with each other
- There should not be too many spots (by ink, mould, etc.) overlapping the track or in its immediate surroundings;
- There should not be spurious indelible signs (like stamps and pencil marks) across the track, which may baffle the detecting algorithm.

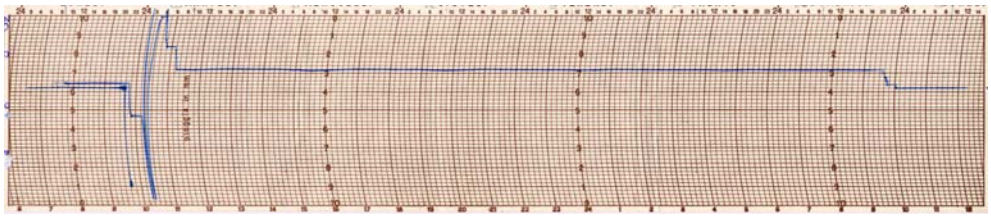


A pluviogram marred by a big ink blot, successfully converted by DigiGraph (Ceccaccio, LT, 07-10-91)

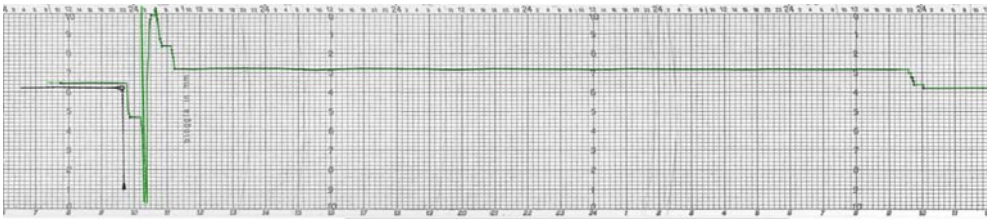


See above: the reconstructed track (in green) does not contain any part of the big ink blot

Sometimes the software makes miracles*:*

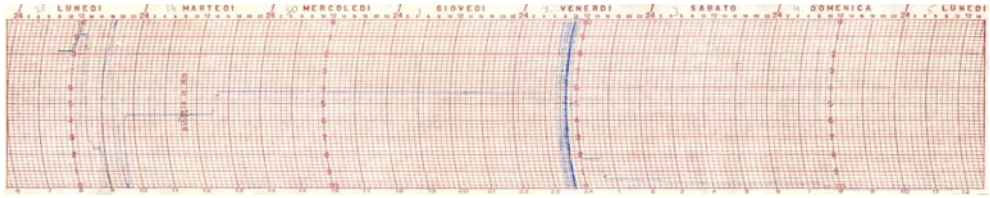


Pluviogram with a double track, a small ink blot sharp peaks, due to an intense rain event, and a systematic offset of the nib, successfully converted (see below)



In the reconstructed track the secondary track and the ink blot are filtered out, the sharp peak is well detected and the offset of the nib has been remedied

Other times the software fails, the main cause of failure for pluviograms, as in the example below, being the saturation of the pluviometer in extreme events which produce densely overlapping track segments. For cartograms of temperature, humidity and pressure, the failure may only be caused by a very bad conservation state of the paper supports.



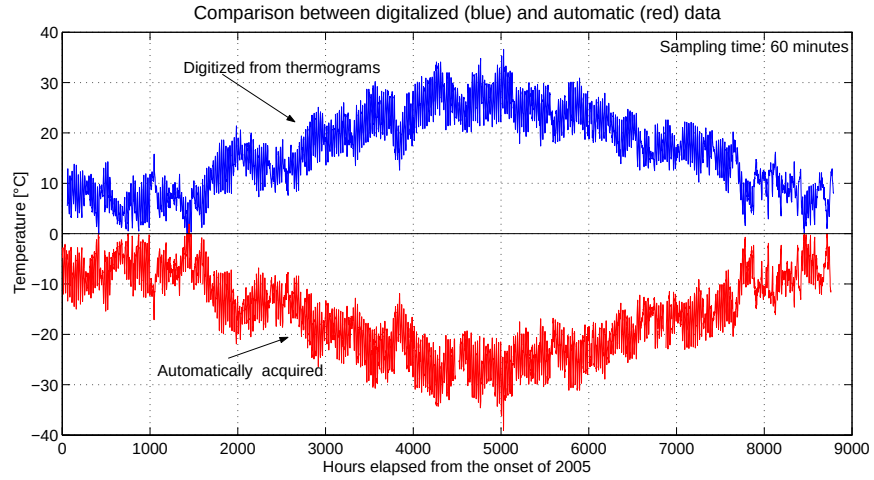
A pluviogram impossible to convert without further manual aid by the user

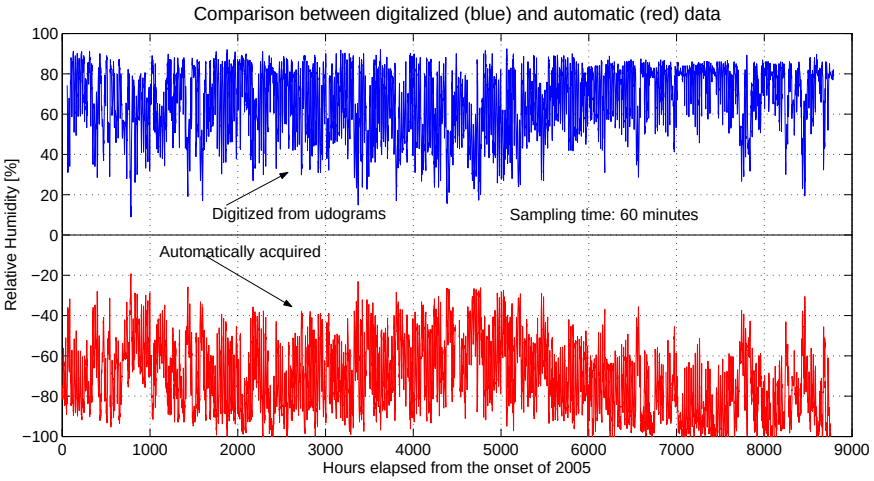
Processing times. If a good quality A3-size scanner is used, the average time for the acquisition and the storage of an image and for its preparation to the digitization process (namely, the compilation of the corresponding configuration file) is less than 2 minutes. Running the program on a high-performance PC, the average execution time results to be near 1 minute with maximum processing times less than 2 minutes.

A case study: a year of data of temperature, humidity and precipitation recorded at the Rome Observatory in Collegio Romano

The station of Collegio Romano, with its tower Calandrelli reaching 66.4 m in the centre of Rome,, is one of the most ancient Italian meteorological observatories, still active today since 1788. For more than 200 years the observations have been mostly done with traditional mechanical recorders based on inked nibs leaving continuous tracks on a graph paper rolled over a slowly rotating drum. Toward the end of last century, this traditional instrumentation has been flanked with modern computer-based automatic acquisition systems, so that it has become possible to make a comparison between data written on cartograms and data stored on files. Of course, the data written on cartograms must be digitized in

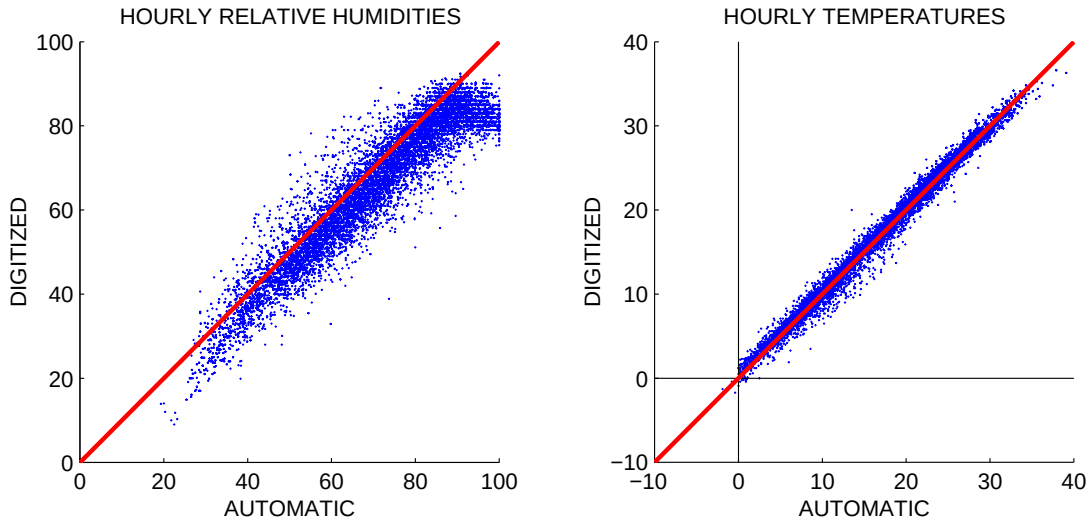
advance at a sufficiently high sampling rate in order that a detailed and exhaustive comparison be possible. This is exactly what the software *DigiGraph* is able to do rapidly and accurately. The whole year 2005 was chosen for the comparison: 52 thermograms, 52 udograms, 52 pluviograms were submitted to *DigiGraph* obtaining observations every 5 minutes interrupted by a brief gap every Mondays, corresponding to the time taken by the manual procedure necessary to replace the paper form on the rotating drum. On the other hand, some data files created by the automatic acquisition system have been processed in order to extract the observations for temperature, relative humidity and cumulated precipitation for the same period. The latter data are available at an hourly sampling rate, except cumulated precipitation that is also available at a faster sampling rate of 10 minutes. Therefore, the digitized data were either re-cumulated every 10 minutes (precipitation) or averaged (temperature and humidity) at each exact hour with a moving average scheme using unequal weights and at most 2 neighbours on each side. This preparation yielded two data-bases for each quantity under examination, that can be easily plotted and compared. The results are shown for the whole year in the following 2 figures, for temperature and relative humidity (precipitation is shown in the next page). The automatic data have been plotted in red upside down (the temperatures have been simply changed by sign) for a better parallel presentation of the two hopefully twin time series. Now, apart from the *prima facie* specular agreement of the two plots (at least on the large time scale), we actually get very different results for temperature and relative humidity.



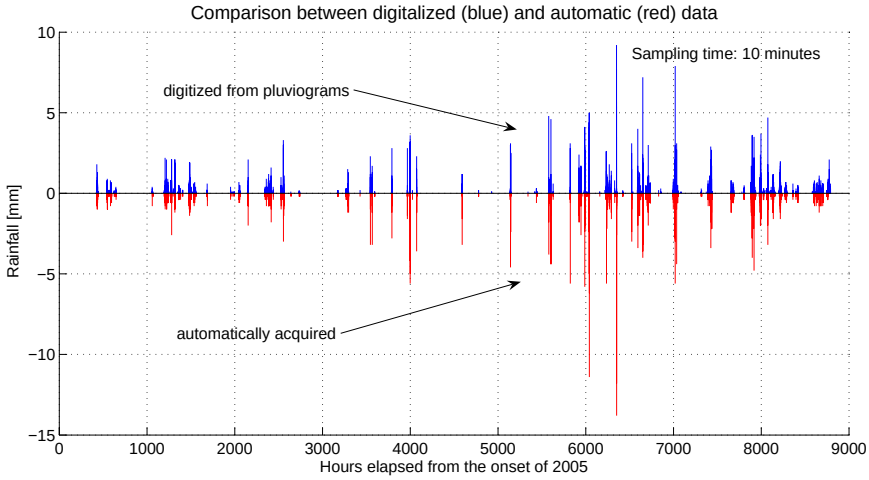


The difference is well illustrated in the following two plots where the ideal situation is that all points should lie on the straight line bisecting the first quadrant. Of course, in both plots only points having the same instant of observation have been drawn. Now, while the points align pretty well along the bisector for temperature, the comparison between automatic and digitized relative humidities evidences two sources of systematic errors in the instrument used to write the udograms. First, there is a persistent offset of about 5 percent units to be added to the digitized values in order to set them in line with the automatic data; second, there is a sudden nonlinear distortion in the upper range of humidities, where the response of the instrument

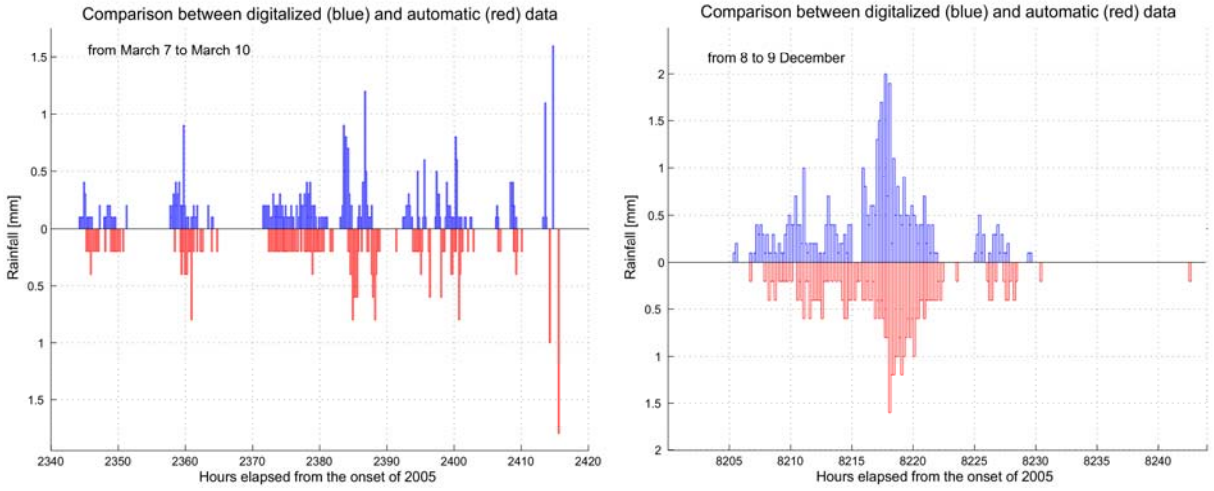
that generated the udograms appears to come early to saturation for values between 85 and 90% . Actually, it is seen from the last figure above that digitized values never exceed 90% throughout the year.



For what concerns precipitation, the following figure shows the precipitation cumulated every 10 minutes in the form of narrows bars, each insisting on the corresponding 10-minute time span. Again, at a first glance, one gets the impression of an overall agreement between the shapes of these two time series, intense events appearing to occur at the same time locations and almost with the same intensities.



However, a closer inspection shows some more or less systematic differences between the two records, as shown in the following two blowups representing in detail stretches of a few days.



The most common feature is an almost systematic time lag of 1 or 2 hours between the digitized and the automatic data with the first ones lagging behind the second ones. Note that this lag can be ascribed neither to the legal time 1-hour shift (that is always ignored either in automatic data acquisition or when positioning the graph paper over the rotating drum) nor to the 1-hour difference between UTC time and Roman local solar time, a difference that has been already taken into consideration (automatic data acquisition happens to use Greenwich Meridian Time). The only source for this frequent difference can be the non exact positioning of the graph paper on the rotating drum, which, in the case of paper forms lasting a week, can easily produce a spurioius time shift of 1-2 hours. It is not completely clear, however, why this misalignment should produce a difference that goes systematically in one direction.

CONCLUSIONS

The software *DigiGraph*, prepared and tested at the Institute of Sciences of Atmosphere and Climate, CNR, Rome, is an easy-to-use application devised for the completely automatic reading and digitization of thermograms, udograms, pluviograms and barograms acquired by means of paper recording instrumentations. The tracks impressed on graph paper, once transformed into image files with the use of a scanner, are converted rapidly and accurately into numeric data files of a format chosen by the user in less than 2 minutes even adopting sampling times as low as 5 minutes.

The use of this software makes thus possible to recover within reasonable times the vast information stored in the voluminous paper archives existing in many organisations and institutions, that have accumulated huge quantities of cartograms throughout many years and in some cases for more than a century, whose management and conservation is itself a serious problem.

As shown in the preceding sections, the software *DigiGraph* has the capability of eliminating automatically many of the various kinds of artifacts that often mar the cartograms, of avoiding the errors due to subjective judgment and to the misreading of curvilinear and nonlinear scales of the human operators. Furthermore, its speed of execution allows for a fast exploitation of the data acquired with the traditional methods at meteorological stations making it possible a rapid comparison between the traditional data and the data acquired with computer-based automatic acquisition systems.

Even when this comparison indicates a conflict, the utility of the fast data conversion is obvious in evidencing, for example, either a systematic bias or a malfunctioning in some traditional instrument that needs a recalibration, or else an inaccurate procedure followed by the human operators.

Since the algorithms underlying the software are enough versatile, it is not difficult to foresee future extensions of *DigiGraph* in order to make it apt to cope with other kinds of graph paper diagrams, like those coming from the continuous recordings of wind intensity and direction.