A general purpose procedure for locating outliers in multivariate time series: Application to an hourly wind dataset

CARLOS LÓPEZ† and ELÍAS KAPLAN‡
Centro de Cálculo, Faculty of Engineering (11)
CC 30, Montevideo, Uruguay
†internet: carlos@fing.edu.uy
‡internet: elias@fing.edu.uy

Abstract: The techniques employed in the treatment of an hourly surface wind database during the development and calibration phases of an objective wind field interpolator model are presented. The model itself has been applied to estimate the regional wind energy resource creating a layer in a GIS environment.
Any model is affected to some extent by both random and sistematic errors (outliers) in the input data. So it is advisable to remove them prior to use the data bank, while keeping at lowest the required effort.
For this case, some different methodologies have been applied. The most succesfull was based in Principal Component Analysis (PCA). It was able to locate outliers with an associated type I and II errors of 49.16 per cent and 6.44 per cent , respectively, in a single step.
The methodology is liable to be used in real time, involving minimum computer resources. For the stages described here, only errors coming from manually digitizing are considered. However, it is suggested that PCA may help in detecting random errors from the observer himself, and also some kind of sistematic errors, all of which is still in an investigation phase.

1. Introduction

1.1 *Presentation of the problem*

In all experimental data banks two sources of errors exist: those inherent to the measuring operation, and those generated in the time of data keying or processing. Both types of error could have an effect more or less important depending to the problem in study. According to Husain, 1989, "..*The failure of many capital intensive projects througout the world can be attributed in part to an inadequate record length, the sparseness of the network, or the inaccuracy of the information*..". In our problem, the wind energy resource proved to be robust against outliers because the hourly values are simply averaged over time. However, this might not be the case for time-dependent models, like evolving atmospheric pollution. There the errors spread in the time, and depending on the characteristics of the problem itself, their effect is more or less persistent and significant.

For some of these models (in the daily operation) it is easy for the user to note important errors, since he is evaluating the kindness of the prediction in the following day or hour.But in the empirical parameters calibration stage it is not

possible to analyze manually a sequence of thousands of measured vs. calculated values. Typical procedures rely on hypothesis about the distribution of their difference and use simple estimators like the standard deviation as an attempt to locate unusual differences.

Such procedure might help in locating events clearly erroneous, but it is unable to point out other more subtles, affecting the value of the automatically adjusted parameters in an uncontrolled way. In order to debugg the data bank, the data in paper taken by the observer have been assumed as error free, and counting as errors only those arising in the process of keying.

1.2 *Methodological background*

For the location of anomalous data, the only national registered antecedent consists in the recommendations developed by the Climatologic and Documentation Office of the National Weather Service (DNM, 1988). The rules for wind are typically a check for acceptable range, both for direction and velocity. Also some simple independent temporal and spatial checks are sketched.

With concerning the random errors, the trend is to compare the measurements with a model of the phenomenon (Francis, 1986; Hollingsworth *et al.* 1986, etc.). The last one asseverates that for the case of the wind, the differences between observations and predictions have a normal distribution approximately. In that case, it is relatively easy to detect the anomalous data and separate them for a later analysis. As a disadvantage, it should be pointed the important volume of information required, as well as the high computational costs involved in creating and operating a model.

If you are unable or do not want to exploit the underlying physics that connect the variables, the pure statistical methods are an alternative to evaluate. Barnett *et al.* 1984 summarizes the different applicable techniques for tackling this problem. In the case of the multivariate data analysis, it can be distinguish two main methodological lines, depending if the distribution function is known or not.The first group includes the so called Discordance Tests, a set of techniques strongly based on hypothesis about the distribution of the sampled data and which requires prior knowledge or estimation of the distribution parameters. Antecedents also exist tied to the case in that the theoretical distribution responds to a type of law and the sampled data to another, as in the case reported by O'Hagan, 1990, where the fact that one of the distributions is normal and the other is of Student´s type enables the use of certain methodology in order to put the anomalous data in evidence.

The second group identified by Barnett are the Informal Methods. They disregard the formal aspects of the data distribution, and aim to exploit other

properties. This group includes among others: a)univariate marginal methods, which derives from the sample a valid range; b) graphic methods, based on looking for isolated points lying far from the data cloud; c) the application of methods of correlation (Gnanadesikan *et al.* 1972); d) the search of generalized representative distances, e) techniques related with cluster analysis (see for example, Fernau *et al.* 1990) and principal components analysis (PCA) (Hawkins, 1974; López et al. 1994a, etc.), among others.

2 The Problem

Since 1988 our team was involved in evaluating the National Wind Energy Resource. Although it was not an explicit objective of the project a comprehensive quality control was performed. Such control was carried out both on the data originated routinely at the DNM (National Weather Service) as well as those of the automatic anemometers of the DNE (National Department of Energy) which were the target locations. On the other hand, in order to have a better evaluation of the wind energy resource, it was necessary to complete the time series to the longest available period. The latter is presented in a companion paper. With these goals some algorithms have been implemented in order to detect anomalous data. Even though we have at hands a model of the phenomena, we preferred an statistical approach.

Five stations were selected for the test, all located to the south of the country, (see fig. 1): Melo (440), Paso de los Toros (460), Treinta y Tres (500), Carrasco (580), Punta del Este (595). They were chosen due to its geographical localization around the automatic stations of the DNE. This also conditions the periods to work with, including part of the years 1990-1991 and the year 1984.

The work carried out consisted then in picking dates (and data for such dates) which behave unusually with respect to the population, then going to the original paper records at the DNM and check there against the files in paper. The value is qualified as erroneous only when the registration in paper doesn't coincide with the magnetic registration available. The process continues with a new step and consultation in the file in paper. Up to eight successive steps were carried out for the same period.

It must be clear that we do not excluded the possible existence of another sources of errors, occurred so much in the process of capture of the data or their transcription to the paper. Some other possibilites are: a) Inadequate exposition of the anemometer to the surroundings, b) lacking in maintenance of the instrument, flaws, etc. c) bad habit in the methodology from taking of measuring, d) physical characteristics of the instrument (speed threshold, characteristic length, etc.). Such problems (which probably exist to some extent in all the stations) are of difficult

correction, in the sense that although they could be recognized after an inspection, the original value has already been lost.

The typing error is the only one that could be documented and corrected, and therefore the indexes that will be introduced should be evaluated keeping in mind that they might also be detecting other errors which exist before the transcription to the paper. The efficiency of the method would be yet better.

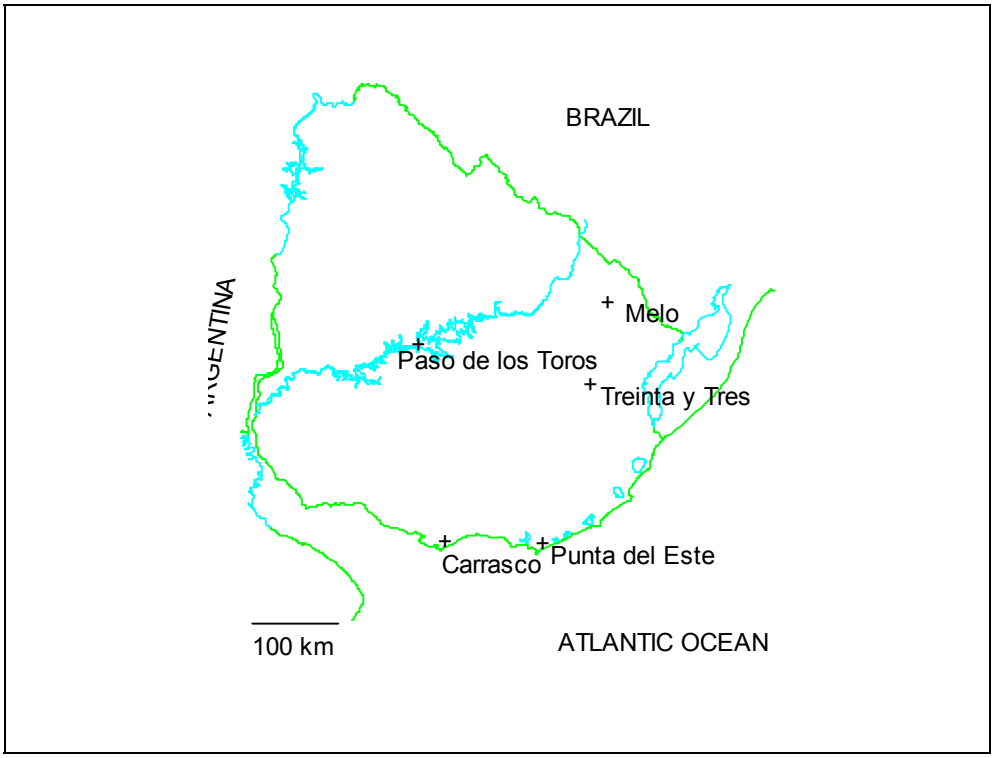

Figure 1 Location of the weather stations

3 The test procedure

The methods were presented in full in López *et al.* 1994a and 1994b applied to the case of pluviometric data (daily values). Here only a brief synthesis will be introduced. It will be referred as event the set of data values for a particular hour. As indicated in López *et al.* 1994b, the population could be divided in a) hours with data in all the considerate stations and b) hours with any absence in them. The process will require measured data, or estimates for the considerate event, in all the stations. So an imputation for the missing values is required as part of the process.

The first stage of the method requires performing a Principal Components Analysis (PCA) (see for example, Lebart *et al.* 1977) While considering events without missing values, it is straightforward to find the principal components of the population. They are a property of the group of events, and not of any event in particular.

Any event has associated 2n measurements (components u and v in each one of the n considerate stations), and it can be sought as a vector in the space R^{2n}. The principal components are a base of that space and we named after scores the projection of the space in that base. The scores are also 2n numbers. It is equivalent to manage the temporal series of the original measurements or the temporal series of the scores, there being between both series a mere lineal transformation.

The time series of the scores are mutually uncorrelated, an important difference with the original values. Such fact enables us to consider each component separately, knowing that there is no redundant information in the others. In the figures 2 to 4 some of the distributions of the observed scores are introduced.

It should be noticed that except for the first three components ("the most important") the observed distribution is relatively concentrated around the zero, as it was pointed out in López, 1993. As indicated in López *et al.* 1994a, it is possible to identify anomalous events comparing all or some of the scores of a particular hour with its distribution in all the population. For each score's probability function distribution, its percentiles 5 and 95 per cent can be determined, giving an objective criteria to classify a particular event as marginal. So, if for an event, the j-th score is marginal, it is considered that in that event there is something abnormal, and it should consequently be checked.

In Silveira *et al.* 1991 some cases of "abnormal" events detected by the procedure described before were individually analyzed, for the case of rain. One could be noticed that this criterion not only detects errors, but rather also they mark some atypicall events , like heavy convective rain episodes very concentrated in the space. Even though in that cases it was verified that they were not errors, it doesn't contradict that they were abnormal events.

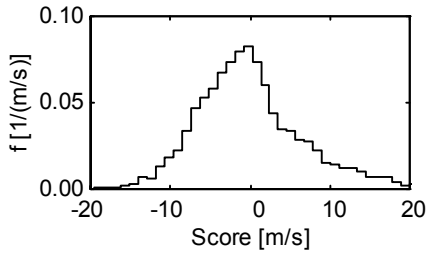

Figure 2 Sketch of some properties of the scores 1 (left) and 2 (right). On top the probability density function; on the middle the power spectrum and on bottom the self correlation against the lag.

Not all the j-th scores should be passed for the criterion. López *et al.* 1994b justifies that there should be an optimal q value, that makes controlling only from the q-th score through the 2n the best option. Such q could only be determined by means of experiments like the one that will be introduced later.

As it was explained, it is possible to detect the anomalous *events* and identify their date and hour. Notice that in the calculation of the scores all the data of that event is involved, so it is not trivial to discriminate which particular station in particular is more likely to have an error. As a solution a sensitivity analysis has been sought using a functional S designed to highlight any unusual situation. S is typically small if all the a_i are themselves not marginal. It is defined as

$$S = \sum_{i \in p} \frac{a_i^2}{w_i} \qquad p = \{k, k+1, .., n\}$$

being a_i the i-th score, and w_i is a weighting coefficient. Hawkins (1974) uses the associated eigenvalues instead of w_i, but we used the criteria suggested in López *et al.* (1994a), which make every term in de summation of the same order. The index i vary within a set p, which in turn depends of an integer parameter k. The S functional neglects the information of the temporal self correlation of the scores. In order to isolate the problematic station it is proposed to calculate for the event in question the partial derivative of the functional $S(u_1, v_1, u_2, v_2, \ldots, u_n, v_n)$

$$\frac{\partial S}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_{i \in p} \frac{a_i^2}{w_i} \qquad j = 1..n$$

where x_j denotes indistinctly the component u or v in the j-th station and being p the set of weighted scores (ranging from 4...10 for example) and a_i is the i-th score. The j that produces the maximum derivative (in absolute value) will identify the most sensitive station, which will be taken as the error candidate. Also the second and third in importance will be taken into account, and they will be qualified as "a", "b" or "c" candidate stations (see Table 1).

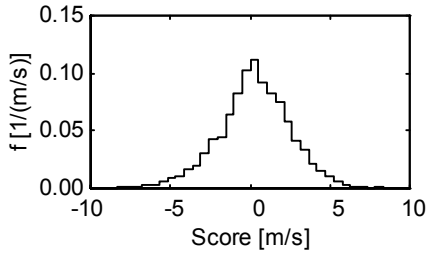

Figure 3 Sketch of some properties of the scores 3 (left) and 4 (right) as presented in fig. 2. Notice the change in scale for the power spectrum. The peak near 0.04 is due to daily variations.

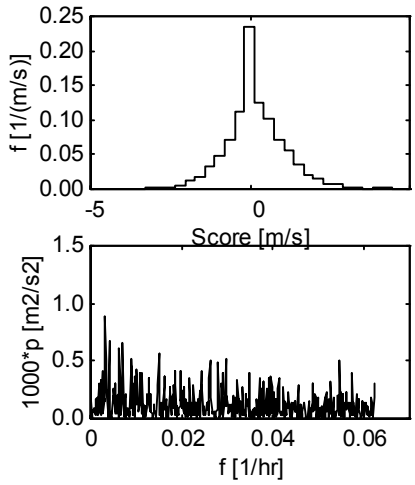

Figure 4 Sketch of some properties of the scores 9 (left) and 10 (right) as presented in fig. 2. Notice again the change in scale for the power spectrum, and also in the x-scale for the pdf.

4 Results

4.1 *Error location on the original data*

In figures 5 and 6 he compared performance of the methods in the case of hourly wind for the year 1984 is presented. As it was explained before, no missing values are allowed for calculating the Principal Components. The available dataset has less than 30 per cent of the events complete, so imputation is mandatory. In order to do so, two different methods were used, named **T**emporal **I**nterpolation of the **P**rincipal **S**cores (TIPS) and **P**enalty of the **P**rincipal **S**cores (POPS). The reader is referred to López *et al.* 1994a, 1994b for further details.

In figure 5 the performance of the TIPS imputation plus error detection is compared against the POPS imputation plus error detection. The original population analized (from year 1984), has 8784 events, rendering 87840 numbers upon multiplying for the 5 stations and keeping in mind that there are two components u and v for each.

The results show that for the TIPS with percentile limits of 0.5 per cent and 99.5 per cent (figure 6, Table 1), 592 events (13.5 per cent of the population) are selected in the first run. Once contrasted with the data "in paper", 301 (50.84 per cent) of the candidates (see Table 1, in boldface) have errors. The rms between the erroneous data and the corrected one is 4.19 m/s. Of these 301 events, 83 (27,57 per cent of the 301) were marked with probability "a" of being the error (column " per cent a_ok", Table 1), 49 (16.28 per cent) with "b" and 41 (13.62 per cent) with "c". A similar table can be devised for the POPS method.

We have also analyzed the intersection of both sets as a separate alternative (i.e., collecting those events which behave atipically after two different imputation methods) but the results are very similar as those presented.

Table 1 Typical results of the application of the TIPS interpolating three of the principal scores. Data from 1984. P indicates step, N indicates new step data, ACU indicates the total accumulated until that step. In boldface values commented in the text.

candidates		% corrected		% a_ok		% b_ok		% c_ok		RMSE	mean
P	N	ACU	N	ACU	N	ACU	N	ACU	N	ACU	[m/s]	[m/s]
1	**592**	**592**	**50.84**	**50.84**	**27.57**	**27.57**	**16.28**	**16.28**	**13.62**	**13.62**	**4.19**	**-0.87**
2	774	1366	28.29	38.07	18.72	23.85	16.44	16.35	10.96	12.50	2.44	-0.47
3	505	1871	17.43	32.50	6.82	21.38	19.32	16.78	9.09	12.01	2.20	-0.50
4	508	2379	16.14	29.00	18.29	21.01	9.76	15.94	20.73	13.04	2.32	-0.10
5	367	2746	31.34	29.32	20.87	20.99	19.13	16.40	14.78	13.29	3.11	-0.00
6	530	3276	7.36	25.76	20.51	20.97	15.38	16.35	15.38	13.39	1.95	-0.41
7	424	**3700**	5.42	**23.43**	21.74	20.99	17.35	15.92	16.12	13.03	2.56	-0.67

The marginal percentile was incremented 0.5 per cent in each step. In the second run said percentile was between 1 per cent and 99 per cent , etc. until the 3.5 per cent and 96.5 per cent of the seventh. This policy allows to select for every step around 500 to 600 new events in order to check with the data "in paper." In the synthetic experiment that will be introduced later, such percentile stayed constant between steps.

The task of checking against the data "in paper" could be regarded as a calibration phase of the method. It is the first stage towards an automatic quality control of incoming or existing data for other years. The results differ greatly as the process goes on. From table 1 can be noticed that the rate of success in the error location of errors varies from 50.84 per cent in the first step, to 5.42 per cent in the seventh one.

By the mere classification as "wrong" - "not wrong" it is not possible to indicate the importance of the detected error. The root mean square (RMS) between the original and corrected value has been used as an estimator. It could be

calculate for each run giving an idea about the incremental effect, or once finished the operation illustrating about the size of the remaining errors.

The process was finished when the procedure found less than 5 per cent of erroneous data among the candidates. The observed incremental RMS was around 2 m/s, implying that the wrong and correct values were too similar. Once checked a total of 3,700 events (containing 37,000 values) the procedure found 867 events with errors in at least one of the stations, that is to say a 23.43 per cent of success in the suggested dates, affecting a 9,9 per cent of the total of events in the year 1984. The associated type I error, defined as the probability of classify as wrong a correct value (Minton, 1969) is estimated as 76.57 per cent . The standard deviation of the discrepancy between the values initially in the files and those on paper resulted 3.5 m/s for the year 1984. If one limit oneself to a single step, the results are a type I error of 49.16 per cent , and an estimated type II value of 6.44 per cent . The type II value is defined as the probability of classify as good a wrong value (Minton, 1969) which in turn require estimate the total number of errors in the dataset. We assumed that we located *all* the errors after the seven steps.

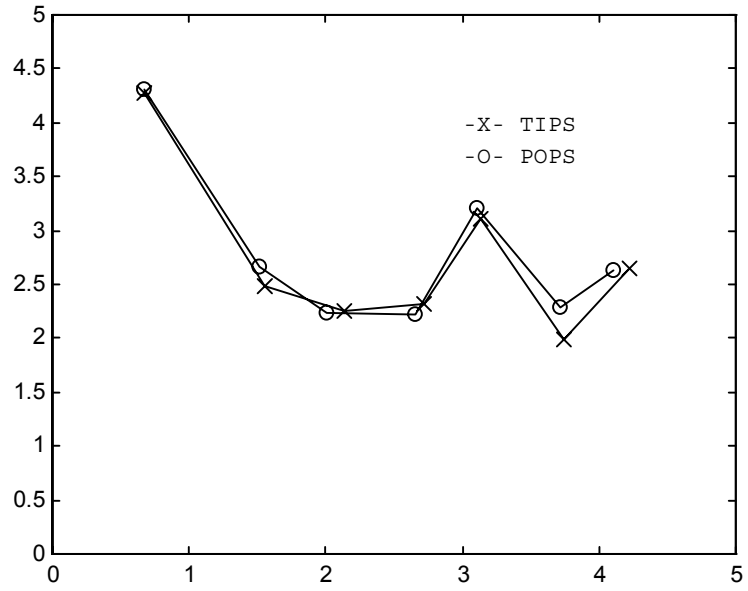

Figure 5 Experimental values of the RMS obtained while depurating the original database. X-axis stands for the fraction of the total database already checked. Y-axis stands for the RMS of difference between wrong values and correct ones on paper.

4.2 *Error detection on the already purified data bank*

The implemented algorithms were evaluated in controlled experiments in order to confirm the kindness of their acting both for detect anomalous data and for imputate missing values in the series.

The experiment consisted in sowing erroneous data and detect them. In an attempt to mimic the behaviour of real errors, it was assigned to the element $v_{i,j}$ of the data table with i and j at random, an element $v_{k,l}$ (multiplied by 2) from the same data table with k and taken l also at random. The cited data table has 10 columns (5 stations for 2 components u and v) and 10171 lines (424 days of the the years 1990-91 for 24 hours) implicating 101710 values as a whole. The factor 2 used with $v_{k,l}$ was used as a crude attempt to resemble the errors observed with real data.

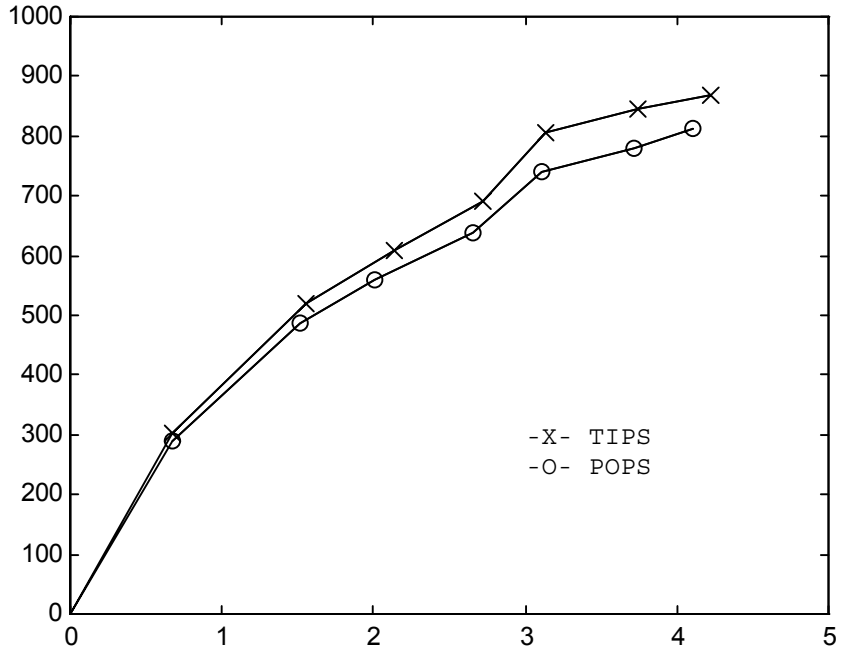

Figure 6 Total number of errors found for a given effort..

They were carried out several tests varying different parameters for the identification of the suspicious data. As an example, in figures 7 to 9 the results with a marginal percentile of 0.5 per cent are shown. We penalized only the prescribed number of terms (those which corresponds to the weakest scores), of the total of 10 scores. The best and robust results of maximal RMS were obtained while using all the terms.

For 13760 erroneous values (13.5 per cent of the total data) the method detected (after imputation with POPS using 1:10 out of 10 scores) 6075 of the artificial errors (44.5 per cent). It was necessary to check 33 per cent of the total, attaining a success rate of 18 per cent (6075 of the 33800) in the revision of candidates.

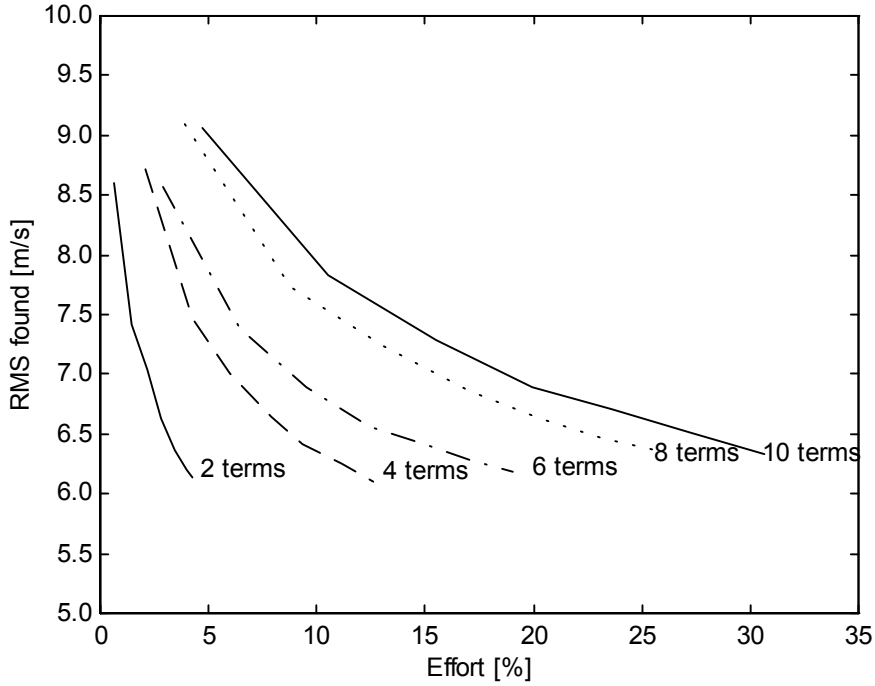

Figure 8 Evolution of the RMS found for the same initial noisy dataset in terms of the effort

Another aspect to keep in mind is the reduction in the remaining standard deviation of the error between original data and erroneous data values. The initial deviation was 4.66 m/s, and after correcting the 6075 erroneous detected values it was reduced to 2.83 m/s. The incremental standard deviation decreases from 9.06 m/s in the first step until 6.24 m/s in the last one. Going further with the calculations for this case (1:10 penalized terms) it can be appreciated that the remaining standard deviation decrease down to 0 (ideal limit that would be attained upon checking the 100 per cent of the data) while the measured incremental deviation stay near 4.66 m/s. The attained value of the remaining standard deviation (2.83 m/s) seems reasonably, since the rms of the database values is 3.98 m/s before correcting them and 3.26 m/s after correcting the erroneous data.

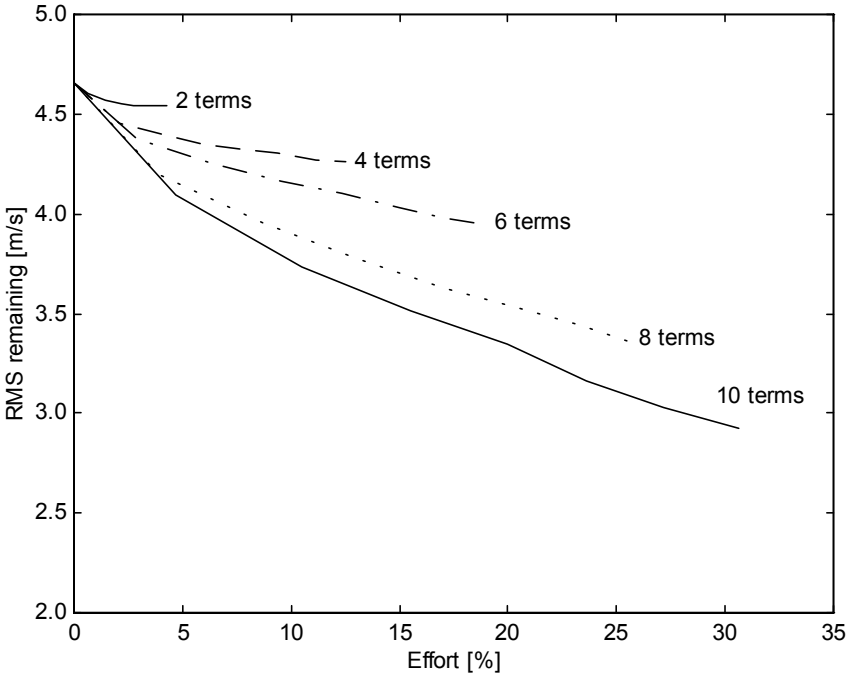

Figure 8 Simulated evolution of the remaining RMS vs. the effort

5 Conclusions

From the performed experiments it could be inferred that:

a) The data should be carefully verified, with the proposed criteria or with another. About 10 per cent of the events have some error.

b) The algorithms employed in order to detect errors performed very satisfactorily.

c) The simulated results in the controlled cases suggest that the real errors could not be simulated by means of the procedure of mixture and multiplication by 2 at random. The performance on real data overcome 23 per cent and the simulated ones 18 per cent .

6 Acknowledgements

It should be recognized the collaboration given by the authorities and personnel of the DNM upon allowing the access to their paper files for the work.

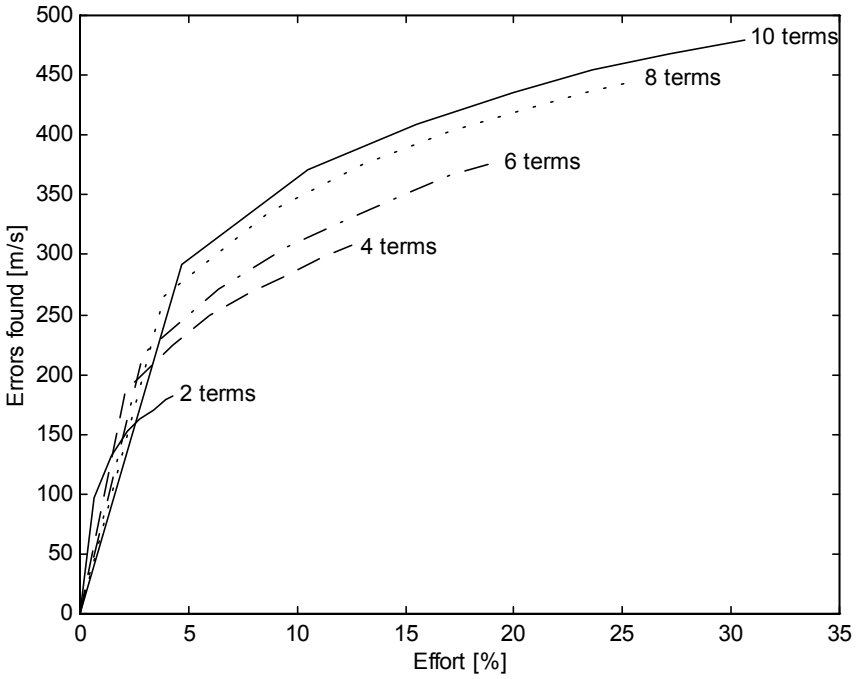

Figure 9 Simulated evolution of the total number of errors found vs. the effort

7 References

Barnett, V.; Lewis, T., 1984. "*Outliers in statistical data*" John Wiley & Sons, 463 pp.

DNM, 1988. "*Procedimientos para el control de calidad climatológico*" Internal Report of the National Weather Service, Nov. 1988, 20 pp. (in spanish)

Fernau, M.E.; Samson, P.J., 1990. "*Use of Cluster analysis to define periods of similar meteorology and precipitation chemistry in eastern North America. Part I: Transport Patterns"* Journal of Applied Meteorology, V 29, N 8, 735-750.

Francis, P.E., 1986. *"The use of numerical wind and wave models to provide areal and temporal extension to instrument calibration and validation of remotely sensed dates*" In Proceedings of workshop on ERS-1 wind and wave calibration, Schliersee, FRG, 2-6 June, 1986 (ESA SP-262, Sept. 1986)

Hawkins, D.M., 1974. "*The detection of errors in multivariate data using Principal Components*" Journal of the American Statistical Association, V 69, 346, 340-344.

Hollingsworth, A.; Shaw, D.B.; Lonnberg, P.; Illari, L.; Arpe, K. and Simmons, A.J., 1986. *"Monitoring of observation and analysis quality by a data assimilation system"* Monthly Weather Review, V 114, N 5, 861-879.

Husain, T., 1989. "*Hydrologic uncertainty measure and network design*" Water Resources Bulletin, V 25, N 3, 527-534.

Lebart, L.; Morineau, A.; Tabard, N. 1977. "*Techniques de la Description Statistique: Méthodes et logiciels pour l'analyse des grands tableaux*". Ed. Dunod, París. 344 pp. (in french)

López, C., 1993. "*Time series forecasting for the windfield over complex topography. Application to southern Uruguay*" M.Sc. thesis, 159 pp., University of Montevideo, Uruguay. (in spanish)

López, C.; González, E.; Goyret, J., 1994a. *"Análisis por componentes principales de datos pluviométricos. a) Aplicación a la detección de datos anómalos"* Estadística (Journal of the Inter-American Statistical Institute) 1994, 46, 146,-147, pp. 25-54.

López, C.; González, J. F.; Curbelo, R., 1994b. *"Análisis por componentes principales de datos pluviométricos. b) Aplicación a la eliminación de ausencias".* Estadística (Journal of the Inter-American Statistical Institute) 1994, 46, 146,-147, pp. 55-83.

López, C.; Kaplan, E., 1997. "*A new technique for imputation of multivariate time series: aplication to an hourly wind dataset*" In preparation.

Minton, G., 1969. *"Inspection and correction error in data processing*" Journal of the American Statistical Association, December, Vol 64, Number 328, pp. 1256-1275

O'Hagan, A., 1990. "*Outliers and credence for location parameter inference*" Journal of the American Statistical Association: Theory and Methods, V 85, N 409, 172-176.

Silveira, L.; López, C.; Genta, J.L.; Curbelo, R.; Anido, C.; Goyret, J.; de los Santos, J.; González, J.; Cabral, A.; Cajelli, A., Curcio, A., 1991. "*Modelo matemático hidrológico de la cuenca del Río Negro*" Informe final. Parte 2, Cap. 4. 83 pp. (in spanish)