KUNGL
TEKNISKA
HÖGSKOLAN

TRITA-GEOFOTO 1997:17
ISSN 1400-3155
ISRN KTH/GEOFOTO/R—97/17-SE

Quality of Geographic Data

Detection of Outliers and Imputation of Missing Values

Dissertation

Carlos López

Royal Institute of Technology
Department of Geodesy and Photogrammetry
Stockholm, Sweden
1997

Abstract

In Geographic Information System (GIS) typical applications data usually comes from a wide range of providers. Such data has variable quality and typically the end user has limited access to the original source (if any). Among other problems those datasets might have missing values and also be affected by outliers. Missing values are common in tabular datasets (like population census, meteorological records, etc.) and the end user is forced to apply any methodology in order to fill the gaps. The data producer cannot recover the missing value and typically does not assign or suggest alternative values. Outliers might arise from careless measurements, instrument malfunction, wrong data processing routines, etc. Current systems give little help to the end user, while the data producer might go back and make another reading, or check the original records if available.

This thesis is concerned with the development and testing of tools intended for two purposes: a) given some dataset, point out dubious values and b) suggest a procedure to assign suitable values for those in doubt or missing. The algorithms were designed in order to be useful for end users as well as data producers.

Only some of the data types usually found in GIS applications have been analyzed, namely tabular categorical data, tabular quantitative data and raster quantitative data. For all of them we suggested new methods and made extensive comparison with traditional alternatives.

For the problem of outlier detection we applied a number of known and new techniques to tabular quantitative data. The examples are from daily precipitation and hourly surface wind records. For raster quantitative datasets we developed and analyzed a new general method suitable for detecting outliers. Digital Elevation Models (DEM) were used as an example. Tabular quantitative (categorical) data (e.g. census data) is also extensively used in GIS applications (opinion polls, economic surveys, etc.). Unfortunately, the procedure cannot be applied to other categorical data typically available in GIS (like a geological or land-use map). For the missing value problem we only treat the case of quantitative tabular data. Most of the methods considered are general purpose, and can be regarded as independent of the dataset. They can be used by the end user as well as the data producer. All the experiment were carried out using MATLAB in UNIX workstations.

KEY WORDS: outliers, blunders, missing values, precipitation, wind, digital elevation models, DEM, categorical data, error model, Geographic Information Systems, GIS.

Acknowledgments

Early stages of this thesis were supervised by *Professor Friedrich Quiel,* and main activities have been completed with the supervision of *Associate Professor Hans Hauska*. His encouragement and support is gratefully acknowledged.

I want to thank also to my present and former colleagues at CeCal (Numerical Analysis Center; Montevideo) for their patience and interest. I should acknowledge also the warm reception at KTH from all fellows.

I wish to thank Professor Roberto Oliveira Mattos for valuable guidance and support. The fruitful collaboration with Prof. Douglas Hawkins was a by side product of the Internet era, and his strong interest is specially acknowledged.

Elías Kaplan introduced me to the secrets of MATLAB. Without his help most of the research could not be finished. Héctor López corrected the english of early versions of the papers.

The research reported here were funded partially by CONICYT (Uruguay) through a scholarship and grants 51/94 and 180/92. Equipment provided through grant ITDC/94-194 by the EU have been used for part of the work. Also partial funding were provided by BITS (Sweden) through a cooperation project with Uruguay.

Despite most of the calculations have been carried out at the Centro de Cálculo, National University at Montevideo, Uruguay, some computer time was provided from ENRIS (mostly *paper IV*) and the Centre for Geoinformatics, Dept. of Geodesy and Photogrammetry (mostly *paper V*) both at the Royal Institute of Technology, Sweden.

Last, but not least, I would like to thank my family, Susana, Carlos Andrés and María Alejandra, for their encouragement during my work on the thesis and for letting me use their time.

List of publications

The thesis is based on the following publications, referred to in the text by their Roman numbers:

I López, C., González, J. F. and Curbelo, R., 1994, "Principal component analysis of pluviometric data a) Application to outlier detection" *English translation of the paper* "Análisis por componentes principales de datos pluviométricos. a) Aplicación a la detección de datos anómalos" *Estadística (Journal of the Inter-American Statistical Institute) 46, 146,-147, pp. 25-54.*

II López, C., González, E. and Goyret, J., 1994, "Principal Component Analysis of pluviometric data b) Application to the missing value problem" *English translation of the paper* "Análisis por componentes principales de datos pluviométricos. b) Aplicación a la eliminación de ausencias" *Estadística (Journal of the Inter-American Statistical Institute) 46, 146,-147, pp. 55-83.*

III López, C., 1997, "Application of ANN to the prediction of missing daily precipitation records, and comparison against linear methodologies" *Third International Conference on Engineering Applications of Neural Networks. Stockholm, 16-18 June, 337-340.*

IV López, C., 1997, "Locating some types of random errors in Digital Terrain Models" *International Journal of Geographic Information Science,* ***11****, 7, 677-698*

V López, C., 1997, "On the improving of height accuracy of Digital Elevation Models: a comparison of some error detection procedures" *Sixth Scandinavian Geographic Information Systems Conference, Stockholm, 1-3 June, 85-106*

VI López C., 1996, "Improvements over the duplicate performance method for outlier detection in categorical multivariate surveys" *To appear in Journal of the Italian Statistical Society,* ***5****, 2*

VII López, C. and Kaplan, E., 1997, "A general purpose procedure for locating outliers in multivariate time series: Application to an hourly wind dataset" *To be submitted*

VIII López, C. and Kaplan, E., 1997, "A new technique for imputation of multivariate time series: application to an hourly wind dataset" *To be submitted*

IX López, C., 1997, "An error model for daily rain records" *Submitted to the Bulletin of the American Meteorological Society*

The research described in papers I, II, VII and *VIII was initiated and carried out by López, who is also the principal author. The co-author(s) participated in the programming and also made contributions to the final manuscript.*

Table of contents

1 Introduction

1.1 Motivation for the project: data and geographic information systems

This thesis is concerned with automated methods for dealing with noisy and/or incomplete databases in Geographic Information Systems (GIS) environments.

Gandin (1988) classified errors in general into two categories: random errors and systematic errors. Random errors are inherent in all data. They are caused by many factors, first of all by the fact that data describe the behavior of the instrument itself, not of what it is intended to measure. Every instrument is approximate by its very nature. Variations in some other parameters, influencing the instrument, may also cause random errors in the measured value of the parameter in question. As opposed to random errors, systematic errors are distributed asymmetrically with respect to zero; their mean values (usually called biases) differ significantly from zero. There are two main causes of systematic errors: a scale shift of the instrument, and an influence of some more or less persistent factor which is not accounted for (or accounted for imprecisely). Systematic errors usually persist in time. This property often allows the determination of even small systematic errors by the application of some Quality Control method to time averaged data. There is a third group: the so-called rough (or large) errors. These are caused by the malfunction of measuring devices and by mistakes during data processing, transmission and reception. Only a very small part of all data is distorted by rough errors. However, these distortions may be very large and may therefore significantly damage analyzed and predicted outputs from models. Very large rough errors, easily detectable at an early stage of the data processing, are usually referred to as "gross errors". Not so large errors might have a chance to remain unnoticed, and may produce significant distortions as well. They are denoted as outliers, which can be defined as "observations that do not follow the pattern of the majority of the data" (Rousseeuw and van Zomeren 1990). Regarding outliers, it is intended to compare different method to locate them in some categories of datasets usually handled by GIS.

The following definition for GIS is due to Johnsson (1994). GIS are computer systems to store, analyze and present geographically referenced information, such as digital maps or point measurements. The information is stored in themes or data layers. Using the analysis tools in a GIS, new information can be extracted by manipulation and combination of existing data layers. The data layers are commonly stored either in raster format (as values in a regular grid, pixels) or in vector format (as points, lines and polygons with associated attributes).

Access to GIS software is also very easy, either using freeware options (GRASS) or commercial ones. Geographic Information Systems is one of the fastest growing markets in software today (Anon 1994). That implies that more people have access to proper tools, and then are able to manipulate and produce data. Data availability will be assured in the future, through the operation of the so called *Clearinghouses*, which will distribute existing datasets to government, industry and the general public (Nebert 1995, 1996).

The combination of widespread data and ready made, easy to use software raises some critical points. John (1993) stated that "...very wrong answers can be derived using perfectly logical GIS analysis techniques, if the users are not aware of the particular peculiarities of data...". For example Openshaw (1989) states that with manual cartographic methods many of the problems associated with map accuracy are visible, and the highly skilled operator makes the necessary adjustments and knows how far the information can be relied upon. With spatial databases (handled by GIS) the equivalent operations are transparent, the operators are no longer so knowledgeable in or aware of the limitations of the data, and the problems are more or less invisible. Despite the fact that Openshaw (1989) focuses mainly on map production, present GIS capabilities allow end users to use sophisticated numerical models and algorithms which take advantage of increasing availability of data to produce answers to complex questions. The problem is that all that output relies on a myth: data is accurate.

This thesis is concerned with part of that problem. Current research is actively involved in describing the accuracy (Goodchild and Hunter 1997; Hunter and Beard 1992; Thapa and Bossler 1992, etc.) and quality of datasets (Buttenfield 1993), while we attempt to design and test tools to improve the accuracy of existing datasets. In order to do so we have selected some typically used datasets, review the literature about error detection, and make comparison tests with them. We prefer general rather than specific-to-the-variable methods, and we will present results on three cases: quantitative tabular, quantitative raster and qualitative tabular data.

For the specific case of the quantitative tabular datasets we also attempted to suggest a solution for the missing value problem, typical for many applications. In each case we discuss the different situations faced by end users and data producers; all the tested algorithms will be of use for both.

The final result of this work is a comparative analysis of a collection of algorithms suitable to be applied within a GIS environment. All of them were tested using real world data.

1.2 The modellers problem in a GIS environment

The typical situation is that someone (*the modeller)* needs to give answers using a GIS system. In order to solve his problem he can devise or use already known relationships between input data and requested results. Such relationship between input and output is named a *model*.

Some problems are:

- imperfect understanding of the relationships (not all the relevant data is used; the relationships are not linear as assumed, etc.). As a result, the model is not accurate enough.
- the lack of available data to test the model make the general validity of the model for other cases dubious.
- given the model, it is sometimes difficult to assess if there is enough data, or if the result depend strongly on the quality of the dataset. Of course it depends!, but the question is how *strongly* it depends.

√ even if an analysis shows that the output of the model is very sensitive to errors, only a few tools to find them exist.

√ even if there is a tool to find outliers, the model might still require some estimates instead of the wrong values. At least a feasible value should be estimated from the dataset.

√ the model (or the implementation of the model) assumes some regularities in the dataset, and might not tolerate missing values.

In order to contribute to the solution of some of the problems denoted with the √ symbol, we attempted to develop tools or algorithms to:

- locate errors for some data types typically used in GIS.
- assign suitable values once a dubious one is found, or conversely, assign values where they are missing.

To situate ourselves in the problem, we present in figure 1.1 a preliminary classification of typical data used in GIS applications. Data can be divided into four categories, and some examples are given on the right. Those examples with a special font are the ones which we have already made some research on.

Once a new method for locating outliers has been developed, it should be compared with other existing methods. To do so, the easiest strategy is to test a single database (Eskridge *et al.* 1995) with known errors previously detected.

Better methods should detect most of known errors, and hopefully some new ones. A Monte Carlo experiment is a natural alternative in order to make a reliable comparison between methods. A tool capable of generating different realizations of datasets with outliers is requested because a single noisy dataset is not enough. There are very few published solutions for this problem. Amrhein and Griffith (1987) and Keefer *et al.* (1988) studied the error location of linear features. Goodchild *et al.* (1992) reported a model capable of dealing with realizations of raster categorical data.

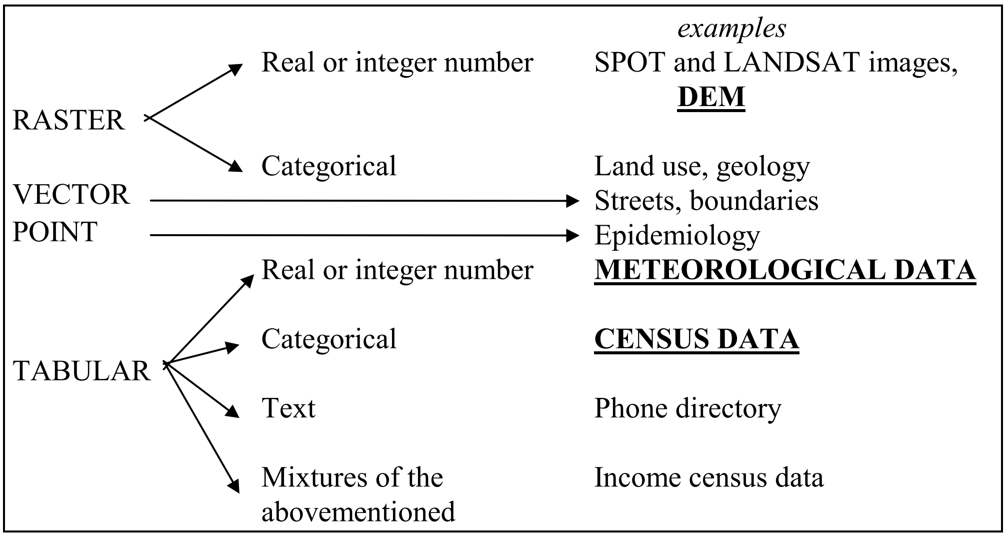

Figure 1.1 A possible classification of data. In bold those considered in this thesis

1.3 Relevance of the investigations

Much of the available data simply do not have any statement about its accuracy, so a natural consequence is that the user almost always assumes that the available data is error free. Even though such hypothesis seems doubtful, none of the currently available GIS software (either commercial or freeware) have standard tools to pinpoint at least the worst errors. The same problem is faced by the data-producer. Also present time GIS software lacks from any built-in capabilities to provide the user with an estimation of the sensitivity of the result in relation with the accuracy of the input.

The situation is also challenging for data producers. Standards for quality are being developed, and they will require to attach information regarding accuracy in the datasets. In addition new technologies and lower costs for both hardware and software make possible to find more sources for a given dataset, leading to more competence among providers. Data collection is still the most expensive part of any spatial information project, and any mistake/error should be detected and corrected in early stages, and with minimum cost. Our conclusion is that the data producer exposes himself to stay out of the market if he is unable to improve the accuracy of his product in a safe and cheap way.

This thesis summarizes some work regarding quality control of a variety of data types. An interesting result is that the same algorithm has been successfully applied to quantitative raster, qualitative tabular and quantitative tabular datasets. Extensive comparisons have been made with other methods for quantitative tabular datasets. The missing value problem has also been considered. The daily precipitation (quantitative tabular) dataset study have been motivated by early developments of an hydrological model for hydropower dam operations. The case of hourly surface wind (quantitative tabular) data was considered while developing a wind energy atlas Both these efforts have been funded by the National Electric Company of Uruguay. The census (qualitative categorical) data example was carried out in the course of the preparatory tasks for the national census of 1996 (Uruguay).

1.4 Objectives

The first objective of this thesis has been to develop and test automatic methods suitable to help GIS data users and data producers to find as many errors as possible in their datasets at minimum effort. Only some categories were considered, namely quantitative and qualitative tabular data, and raster quantitative data. In order to make our results as general as possible, we disregard methods specific to the variable. For example, we did ignore that the wind field should be mass conservative, and we apply to it the same schema used for the daily precipitation. This topics were covered in *papers I* and *VII* for the case of tabular quantitative data (daily precipitation and hourly surface wind records respectively); in *papers V* and *IV* for the case of quantitative raster data (digital elevation models), and in *paper IV* for the case of qualitative tabular data (housing characteristics). For the meteorological data case and one of the DEM papers the methods were also used to find real errors existing in the database. For the categorical data example the test was applied to detect only simulated errors.

The second objective was to develop and compare algorithms for eliminating missing values to the case of quantitative tabular dataset. It has been covered in *papers II and VIII* for the case of precipitation and wind respectively.

1.5 Organization of the thesis

The thesis is organized into six sections. After this first introductory section, the background for the approach taken in the thesis is presented (section 2). Section 3 covers data, study areas, systems and software. The methods used or developed in the conducted research are summarized in section 4, while the results are presented in section 5. The full papers are available in appendices 1-9. Section 6 contains a general discussion of the results, in relation to the issues presented in section 2. The list of references that follows after section 6 refers to the main text. The individual papers contain additional references.

2 Quality, data, models: interaction within a GIS

This section provides the background for the approach to error detection and correction that has been taken in this thesis. Particular emphasis has been put on arguments in favour of methods which are general instead to specific-to-the-variable ones.

2.1 Introduction

This thesis was motivated by the author's early activities developing numerical models, where large datasets of meteorological data were involved. Those datasets were of unknown quality and accuracy and the problems faced using them cannot be considered unique, but a typical example of what is an everyday problem in complex projects. There are at least two aspects to consider: a) the sensitivity of output of the mathematical model in terms of the input data and b) the accuracy of the data. If the model output is insensitive (or *robust*) to the input data, and if the accuracy is not too low, the user can be confident about the results. In other cases a question mark should be included in the results. In this thesis we will present some results concerning the aspect b) and its closely related counterpart: how to imputate missing values.

2.2 Some problems of modeller´s task

The relationship between quality of the model, quality of the input data and quality of the results is well known, and certainly before the existence of Geographical Information Systems. However, the present development of GIS applications, based on good interfaces and taking advantage of all hardware improvements poses a great concern about the easiness to produce methodologically correct answers using inappropriate data. John (1993) stated that "*...very wrong answers can be derived using perfectly logical GIS analysis techniques, if the users are not aware of the particular peculiarities of data...*". This situation can be summarized by the well known phrase "*garbage in, garbage out*". Data might be inappropriate due to scale, attribute, etc. as well as accuracy for the particular problem.

We will extend the use of the term model to describe a well defined operation which is composed of a sequence of steps. So, for example, intersection of coverages to obtain a new coverage will be considered a model as well. We will assume that the model can be expressed as a computer program (either deterministic or stochastic) which using data as an input can produce also data as an output. The output could in turn be compared with independent sources. Many models falls into this category, but not all of them. Once the model and its data is available, problems still remain to be considered, and we will discuss some of them in the following paragraphs.

2.2.1 Lack of independent and appropriate data to validate the model

Data availability will be facilitated through the operation of the so called *Clearinghouses*, which will distribute existing datasets to government, industry and the general public (Nebert 1995, 1996). Thapa and Bossler (1992) remarked that when setting up a GIS, most of the costs (maybe up to 80 per cent) are related to acquiring and/or collecting data. Sharing data between agencies will in turn lower costs to end users and the GIS community as a whole, so there are chances that there will be *too much data* in the near future to consider in models. But their accuracy will still be a problem.

2.2.2 Insufficient knowledge of the sensitivity of output to parameters and data outliers

The quality (hopefully!) and complexity of mathematical models will increase as soon as more data (even unreliable one) is available. This should not be a concern by itself. However, few models take into consideration the accuracy of the data, which is one aspect of its quality. Also, present GIS software lacks tools to warn the operator about the reliability of the results produced. Goodchild and Gopal (1989) state: "*...No current GIS warns the user when a map digitized at 1:24.000 is overlaid with one digitized at 1:1.000.000 and the result is plotted at 1:24.000, and no current GIS carries the scale of the source document as an attribute of the dataset. Few even adjust tolerances when scale changes. Most vector systems perform operations such as line intersections, overlay or buffer zone generation at the full precision of the coordinates, without attention to their accuracy. As a result, inaccuracy often comes as a surprise when the results of the GIS analysis are checked against ground truth, or when plans developed using GIS are implemented. An agency proposing a GIS-based plan loses credibility rapidly when its proposals are found to be inconsistent with known geographical truth...We can now produce rubbish faster and with more elegance than ever before...*".

Our opinion is that models can evolve to use accuracy information (provided it exists) in terms of sensitivity analysis of the output in relation to the input. There exist programs capable to create FORTRAN, C or C++ source code which calculate the exact Jacobian of a function defined by other source code (Griewank *et al.* 1996). This approach is well suited for models which encompass complex relationships between the variables, and might help to estimate the uncertainty in the final result provided the function is derivable and the first order approximation is valid (i.e. the uncertainty is *small*). Some attempts have been made to analyze error propagation in some simple GIS operations using Interval Analysis (Moore 1968), but this approach might be too conservative (no consideration is given to compensation of errors) and too complex except for very simple operations.

2.2.3 Where are the outliers that matter?

Provided that we have decided that the problem at hand is sensitive to errors there is a need for a tool for locating them. We are not aware of general tools for error detection within GIS environments. Some tools exist in statistical packages like SAS or Stata, mostly oriented towards tabular quantitative data. To develop new algorithms, and to test or validate existing ones for application in a GIS environment is one of the main objectives of this thesis.

Early contacts with tabular quantitative datasets defined to a great extent the primary strategy for the approach: how to reformulate the problem in order to use methods developed for tabular quantitative datasets. We do not claim that this is an universally valid solution, but we will show in chapter 4 that for the considered cases the reformulations proposed led to good results and even produced fresh material for further investigation. As examples we might mention analyzing the DEM without using spatial autocorrelation, or the interpretation of the synaptic connections within an Artificial Neural Network.

One argument in favour of using those tabular oriented methods is the weak formal requirements on the data itself. For example, this implies that, provided some dataset has a gaussian distribution, it is irrelevant if its entries are precipitation readings, the square root of the precipitation or a difference between the forecasting from a model and the observed values. The algorithm and its code will be valid as well. This fact encourages us to look for and test methods as generic as possible, avoiding using others which rely on some properties unique for the variable. This is in line with the requirements of a GIS user which might deal with different variables for the same dataset type.

In early stages Principal Component Analysis (PCA) would have been considered as the unifying idea of this research. We found that belief too narrow, and instead we prefer to consider this work as an example of how to extend the applicability of well studied, well established methods developed in the statistics community to the emerging, rather young, GIS needs. The case of PCA is only one possible example. However, it was useful in order to show how powerful might be to establish the connection with tabular oriented methods, and use them to different data types.

2.2.4 How to fill the gaps?

Some models simply assume that there are no missing values. Some might cope with them without further damage. The problem of missing values should not be confused with undersampling. In the former we have defined a measurement strategy and due to uncontrolled reasons the values are lost; in the latter (also named *missing by design*) we obtained the data and we realize later that more data would have been required.

The missing value problem can be addressed for many applications in GIS with interpolation techniques (either spatial or temporal). When interpolation is not feasible nor advisable, maybe some mathematical models might forecast or predict the missing value. Assign a suitable value can be regarded as a "dual" version of the quality control; the "best" interpolation method should produce a "typical" value, thus producing an outlier is unlikely. Many quality control procedures rely in that fact to pinpoint to errors, by analyzing the difference with the interpolated value. Another important connection with outlier detection can be devised in order to provide an answer to: what to do after identifying an outlier, if we do not have access to the true value?

2.3 Concluding remarks

The increasingly growing availability of data from multiple sources is to be guaranteed by the availability of specialized services on the Internet, which will host and deliver existing data with nominal or no cost at all. Improvements in hardware and software will spread the interest, the need and the use of both data and existing mathematical models to a non specialized audience, which is not aware about the uncertainty of the results in relation to the input data. So it is believed that data quality is emerging as one of the most important issues in GIS technology for the next years. Its management requires methods to describe, visualize and measure it properly (see Hunter and Goodchild 1996). Standards to describe the quality are presently under development.

Different models using the same data might pose different requirements on its accuracy, which should be clear after a sensitivity analysis. To our knowledge, few if any models warn the user about the ill-conditioned characteristics of the results. In addition, present generation datasets lacks from estimates of the uncertainty, which precludes considering it in most present models. The trend in spatial data will encourage the specification of uncertainty, and hopefully next generation models will be able to use it. This topic will not be further considered here.

In any case, once the dataset is obtained (either for the end user or the data producer himself) further efforts to improve accuracy should be as effective as possible, because data acquisition is still the most expensive part of setting up a

GIS. This thesis reports some results on that subject and provides some examples of how to apply general statistical tools to a sample of typical datasets usually available or required for GIS applications. The examples and methods have been chosen in order to be non-specific to the physical meaning of the data itself.

3 Materials

In this thesis a few different types of data were considered. In this chapter, the study areas are briefly described along with the data, systems and software that were used. The reader is referred to the individual papers for more details.

3.1 Study areas and data

3.1.1 Main characteristics of the daily precipitation dataset used in *papers I and II*

The weather stations used in the work belong to the Río Tacuarembó catchment. It comprises an area of about 20.000 km^2, and its center is at 32°S 55°W in the NE of Uruguay. The typical landscape is smooth, with elevations below 500 m, few canyons and lakes. The typical monthly precipitation is between 74 and 120 mm/month.

Table 3.1 Listing of the available stations for the Tacuarembó river catchment area

N°	*Name*	*Latitude*	*Longitud*	*Elevation ASL*
1224	Paso Ataques	31°12S	56°21'W	180 mts
1301	Paso del medio	31°27'S	55°04'W	
1379	Moirones	31°36'S	55°58'W	195 mts
1405	Tacuarembó	31°42'S	54°58'W	190 mts
1454	Vichadero	31°47'S	54°41'W	190 mts
1537	Pueblo Noblía	31°57'S	54°07'W	220 mts
1565	La Hilera	32°05'S	55°40'W	
1572	Cuchilla Caraguatá	32°07'S	55°54'W	
1617	Paso Mazangano	32°06'S	55°40'W	
1650	Clara	32°13'S	54°43'W	
1653	Paso Laguna	32°15'S	54°25'W	
1658	Paso Aguiar	32°17'S	54°50'W	
1743	Paso Pereira	32°26'S	55°14'W	

For the work we selected a subset of 13 stations, located as shown in fig. 3.1, which have been carefully checked for typing errors by using the algorithms to be presented later. We restricted ourselves to daily records from Jan 1st. 1975 to Dec 2nd. 1989, covering almost 15 years. Readings are taken usually at 7:00 AM by non dedicated operators (mostly Railway Company employees or local police), and submitted to the county police headquarters by radio. Later in the same day the information is collected by phone by the national electric utility company, which uses it for the operation of three important hydropower dams. The observers also fill in a paper form which is collected once a month.

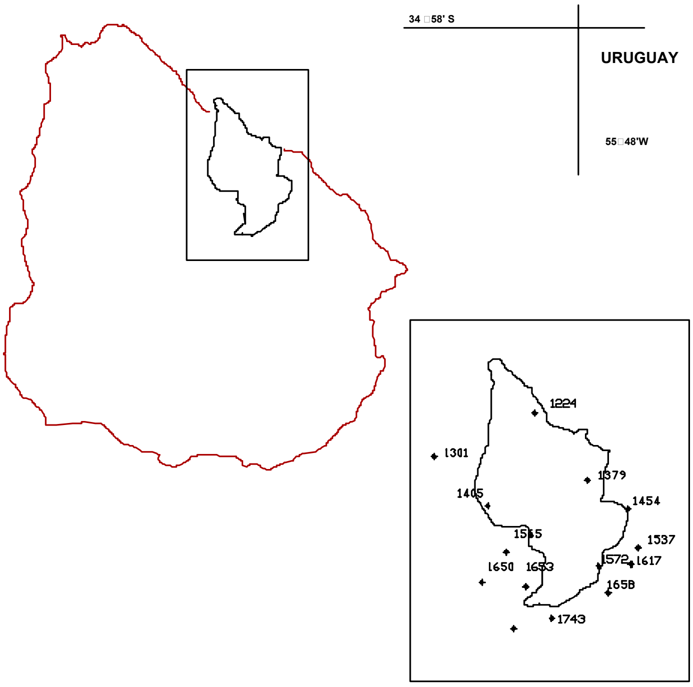

Figure 3.1 Overall location of the pluviometric stations of the Tacuarembó river catchment used in papers I and II

3.1.2 Main characteristics of the daily precipitation dataset used in *paper III*

The Santa Lucía catchment covers an area of 13.600 km^2 and is located in the south of Uruguay, between the 55°W and 57°W and 33°40'S 34°50'S. The yearly precipitation is around 1000 mm/year, with little spatial variation. However, there is a substantial variation in time, with maximum values in 1959 (1600 mm/year) and a minimum in 1916 (500 mm/year). The dryest month is july, with an average precipitation of 75 mm/month, and the wettest is March, with 100 mm/month. The relative humidity ranges from 60 per cent in december (summer) to 78 per cent in june (winter); the annual mean is around 60 per cent. Ten stations belonging to the National Weather Service were chosen (see fig. 3.2) for the analysis. Typical records show that over 80 per cent of the readings are zero; dry days (i.e. zero precipitation in every station) account for more than 30 per cent of the events. The collection procedure is similar to the one in the Tacuarembó catchment, except that the information goes directly to the National Weather Service headquarters in Montevideo.

Table 3.2 Listing of the available stations for the Santa Lucía catchment

N°	*Name*	*Latitude*	*Longitud*	*Elevation ASL*
2436	Puntas de Sauce	33°50'S	57°01'W	120 mts
2486	Pintos	33°54'S	56°50'W	100 mts
2549	Barriga Negra	33°56'S	55°07'W	95 mts
2588	Casupá	34°06'S	55°39'W	124 mts
2662	Cufré	34°13'S	57°07'W	92 mts
2707	Raigón	34°21'S	56°39'W	37 mts
2714	San Ramón	34°18'S	55°58'W	70 mts
2719	Ortiz	34°17'S	55°23'W	115 mts
2816	Joanicó	34°36'S	56°11'W	35 mts
2846	Olmos	34°44'S	55°54'W	40 mts

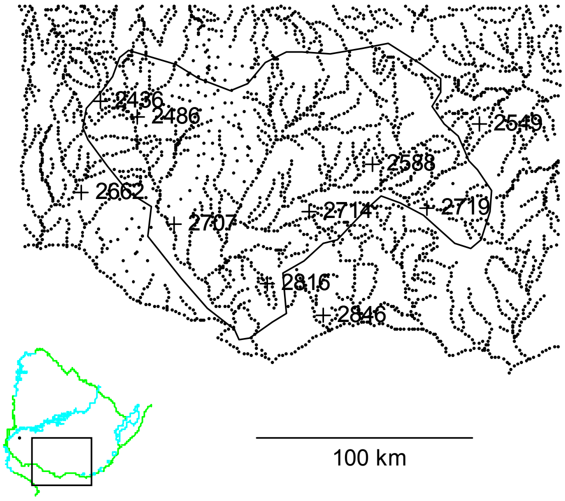

Figure 3.2 Location of the daily precipitation weather stations used in paper III

3.1.3 Main characteristics of the hourly surface wind data used in *papers VII and VIII*

For the experiment described in *papers VII* and *VIII* five stations belonging to the National Weather Service were selected, all located in the south of Uruguay, (see fig. 3.3): Melo, Paso de los Toros, Treinta y Tres, Carrasco and Punta del Este. Both *papers VII* and *VIII* report partial efforts toward the development of a National Wind Energy Atlas.

Table 3.3 Listing of the available stations for the wind energy atlas project (from *paper VII*)

N°	*Name*	*Latitude*	*Longitud*
595	Punta del Este	34°58'S	54°57'W
580	Carrasco	34°50'S	56°00'W
500	Treinta y Tres	33°13'S	55°07'W
460	Paso de los Toros	32°48'S	56°31'W
440	Melo	32°22'S	54°11'W

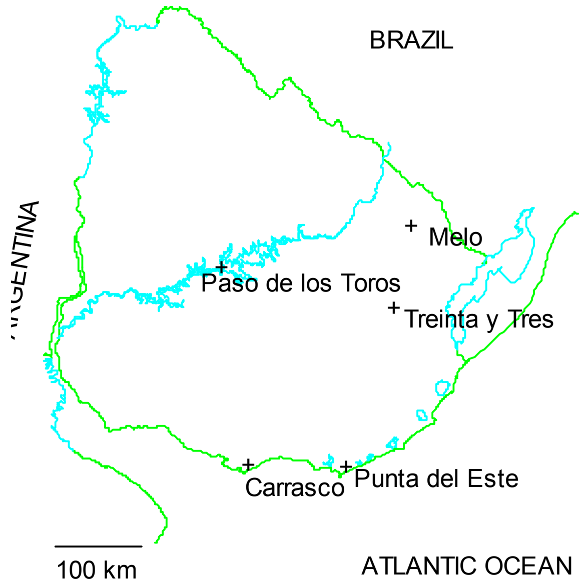

Figure 3.3 Location of the surface wind stations used in papers VII and VIII

Since it was intended to compare the predicted wind with the observed one, some independent instruments were installed in the field for a short period. The choice of the five weather stations were motivated by its geographical localization around the target area. The availability of field records also conditions the periods to work with, including hourly records from part of the years 1990-1991 and the year 1984.

3.1.4 Main characteristics of the DEM datasets used in *papers IV and V*

The DEM used as a test case in *paper IV* covers an area of 7.5x5 km in Stockholm with 150x100 points with a 50 m grid spacing and 1 m elevation resolution. The area consists mainly of hilly terprecipitation, with elevation values ranging from 0 to 59 m. The DEM has a mean elevation value of 20.83 m and its standard deviation is 9.47 m. No data describing quality or accuracy were available.

For *paper V* we have selected two DEM of the Aix-en-Provence region in the South of France. The extent of both is 12.42 km by 6.9 km with 30 m spacing. We used a subset of 360 rows and 216 columns for all calculations. Both DEM include Montagne Sainte Victoire (elevation 1044m). One DEM has been derived from a set of three SPOT images using a stereo matcher (see Day and Muller 1988 for details), and further interpolated to the same grid by using values within a window of size 21 pixels. Elevation values have been kriged using a spheric variogram of 4000 m^2 sill and 3000 m range, assuming an accuracy for the window of 11 m S.D. The other DEM has been produced by manual photogrammetric measurement of spot elevations from contemporaneous underflight aerial photography. Its accuracy has been estimated by multiple set-up and observation of several blocks within the DEM.

3.1.5 Main characteristics of the census dataset used in *paper VI*

A subset of the raw data collected and processed during the National Census of 1985 in Uruguay were used in *paper VI*. It includes data on housing characteristics in the Flores region (located in the center of the country). We considered only private houses cases without missing values. The final set has 4963 records, but to decrease computer time requirements, the simulations were carried out over only 2500 records. The dataset is claimed to be typed twice, which implies that two independent magnetic databases have been created and its discrepancies corrected; the original paper records are not available.

3.2 Systems and software

All the calculations were performed using Matlab, with the exceptions of *paper I and II* and the calculation of the Minimum Volume Ellipsoid (MVE), Minimum Covariance Determinant (MCD), Least Trimmed Squares (LTS) and Least Median of Squares (LMS), where FORTRAN was used. Matlab is a package for technical computing that combines numeric computation, advanced graphics and visualization, together with a high level programming language. Matlab has been chosen due to its suitability for analysis, algorithm prototyping and application development. Matlab's Neural Networks Toolbox have been extensively used, and no attempt to modify or complement it has been done. Most of the statistical routines have been writen from scratch.

The calculations for the work described in the early *papers I* and *II* have been carried out in an IBM 4341 mainframe; the calculations for the other work have been done in the UNIX environment, using either SUN, DEC-ALPHA and BULL equipment (in Uruguay) and SUN and Silicon Graphics (in Sweden).

4 Aim and Methods

In this section the methods developed or applied in the individual studies are summarized. The methods described in section 4.1 are related to quality control of quantitative datasets. In section 4.2 we will analyze the methods applied to the missing value problem. Section 4.3 consider the quality control problem for raster datasets and finally section 4.4 is devoted to the case of the qualitative tabular data.

4.1 Quality control of quantitative tabular data

4.1.1 Aim

While developing a deterministic, conceptual hydrological model for the Río Negro basin in Uruguay using daily precipitation and flow as inputs and predicting from one day to two weeks ahead flow, it was required that many empirical coefficients related with soil properties were adjusted in order to fit available flow data. The robustness of those coefficients against outliers were not known, and the main source of concern was the daily precipitation records, so we designed a method suitable to pinpoint unlikely values. After a direct check against available records on paper, we were able to locate at least those errors arising from typing and further computer processing. All of this was considered in *paper I.* As another example with the same methodology we analyzed hourly surface wind records, which have been used in the development of a Wind Energy Atlas of Uruguay and the author's master's thesis. The results were presented in *paper VII*.

Within the framework of a specific research project we extend our previous results to compare more methods for outlier detection. A Monte Carlo approach together with an outlier-generation mechanism was developed in order to validate the capabilities of the methods.

In early stages of this research we decided to focus on methods which are not rule-based (which in turn require an expert who provides the rules) and also not model-based (which explicitely account for mathematical relationships between variables), because of their lack of generality and the requirements on CPU and extra data. We preferred *general* and *objective* methods; *general* because they are not specific to any variable and *objective* because they can be used independent of the previous knowledge of the operator. The methods have been tested with some datasets believed to be representative of a wider class of variables. In the following we will describe briefly examples of rule based methods.

4.1.2 Rule-based and model-based methods

Weather is routinely forecasted at a global scale using sophisticated software, hardware and communication facilities in very few institutions in the world. In all cases the process of feeding the calculations with the new observations is crucial, and much attention is drawn not to include wrong values into the calculations. Data is collected from all over the world every hour or every six hours, and a quality control is an integral part of the Data Assimilation System. This term applies for the package which receives and processes the data, and produces grid values which will be the input of the numerical models. Quality control algorithms are designed to modify or reject erroneous meteorological data. Some of the checks have been operative from the beginning and they might require temporal and/or spatial continuity (or consistency) with the neighbors. Other checks depends on feasibility rules among different variables (rain without cloud cover) which will flag a record as wrong. An important group using dynamic relations such the geostrophic one (relating the pressure field with surface wind) is also considered there. We want to mention a couple of examples.

Bruce *et al.* (1995) applied both simple temporal tests and spatial tests using some different imputation strategies to monthly precipitation. Every record in the dataset is compared against the best available interpolated one and it is flagged to fail either the temporal, the spatial or both tests if the absolute difference exceeds a prescribed multiple of the interquartile range. The best method is defined after an analysis of the performance for each month; a possible conclusion is that Optimum Interpolation might be preferred for May, and not for June for example. Using a world wide dataset holding 5899 stations for the 1951-1981 period, they claim that about 90 per cent of the *monthly* precipitation records passed both tests.

Another realistic example can be quoted from the work of Reek *et al.* (1992). They reported on the quality control procedures used on an over 100 years long climatic database, collected as a collaborative effort from many sources in the U.S.A. The quality control procedure for this dataset include type all records twice since 1989; routine key punch started in 1962. The quality control reported by Reek is based on rules, and once a record fails to satisfy them, an attempt is made to correct it. If the changes failed to produce a feasible value, the record is submitted for human analysis. The correction heuristics model typical key-entry errors: for example shift in the decimal point, zero reading instead of blanks, a "1" in the first digit of termperature readings (leading to 153°F instead of 53°F) and also wrong sign. Other rules are specific to the variables handled. They used an expert system to collect and analyze the output of all tests and to decide what to do.

4.1.3 Principal Component Analysis based methods (*paper I)*

The procedure proposed for the first time in *paper I* is based upon Principal Component Analysis (PCA). Since it will be used in *all* the forthcoming examples, it is fit to give a brief introduction to it. Any tabular dataset with n events (rows) and w variables (columns) can be represented as a cloud of points in R^w. The correlation between rows is ignored. PCA attempts to find the direction $\mathbf{e_l}$ of the vector in R^w space which minimizes S, defined as the sum of distances M_kH_k squared, taken over all k (fig. 4.1). The origin **O** is the centroid of the set of points. For the sake of clarity, points with negative coordinates are not shown in the figure. The projection OH_k, which is also the scalar product of vector $\mathbf{M_k}$**-O** with the unit vector $\mathbf{e_l}$**,** is called the score (after Richman 1986).

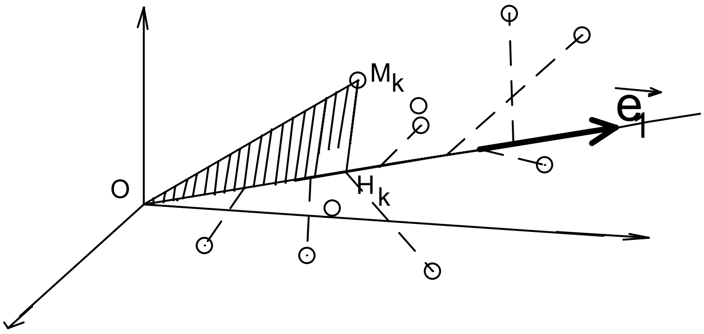

Figure 4.0 Sketch of the first principal component, for w=3 (from paper I)

Thus $\mathbf{M_k}$**-**$\mathbf{H_k}$ is orthogonal to $\mathbf{e_l}$. There is one score value associated with vector $\mathbf{e_l}$ for each point in R^w. Let us also assume that $\mathbf{e_l}$ is unique. If all the values M_kH_k are zero, we have reduced the problem of original dimension w, to a one-dimensional one. All the variability in the observations is explained by a single vector $\mathbf{e_l}$**.** If this is not the case, we may try to repeat the procedure with the remaining variability M_kH_k, which belongs to a (w-1) subspace of R^w orthogonal to $\mathbf{e_l}$**.** The original measurements $\mathbf{M_k}$ **- O** can be replaced with the difference $\mathbf{OM_k}$ **-** $\mathbf{OH_k}$, which is equal to $\mathbf{M_k}$ **-** $\mathbf{H_k}$.

For the new cloud there should be a vector $\mathbf{e_2}$ (orthogonal to $\mathbf{e_1}$) which minimizes the distance S in the R^w space. The process continues until w vectors $\mathbf{e_p}$ have been created; each new vector $\mathbf{e_p}$ being orthogonal to all the previous ones. The vectors $\mathbf{e_p}$ are called principal components (PC). Each event $\mathbf{M_k}$ **- O** can be expressed as a linear combination of the PC's

$$\mathbf{M}_k - \mathbf{O} = a_1(k) * \mathbf{e}_1 + a_2(k) * \mathbf{e}_2 + a_3(k) * \mathbf{e}_3 + \ldots + a_w(k) * \mathbf{e}_w \qquad (1)$$

It can be shown that the scores $a_i(k)$ associated with vector $\mathbf{e_i}$ are uncorrelated with those of vector $\mathbf{e_j}$. The vectors $\mathbf{e_i}$ are the eigenvectors of the covariance matrix of the data, and its components are named *loadings* in the literature. The sum of the corresponding eigenvalues equals the sum of the squares of the distances M_kH_k (Lebart *et al.* 1987).

PCA analysis generates a sequence of principal components, which explains most (or all, for p=w) of the variance of the data. This implies that the RMS of the error in approximating the data with a linear combination of their first p vectors is a minimum for a given p<w; (p=1 in fig. 4.1). It has been shown that in most cases a good approximation of data is achieved for p<<w. Since for p=w the w PC's form a basis in R^w space, they can represent without error any of the n points in the set, using scores as weights. In *paper IV* we claim that some of the scores contains essential information on the structure of the cloud, while others are more related with noise. Once identified, such scores were used to pinpoint those points in R^w space which are prone to hold an error. The outliers are denoted by unusually large values for at least one of those scores. So points with arbitrary values for the structural scores are not regarded as outliers, while a different interval is specified for each noisy score in order to be considered outlier free.

However, find the point is not the complete answer to the problem because each point depends on w independent values, and the point might be outlying due to only one or few wrong values. This situation is common for most methods available in the literature, which might be considered *event oriented,* as opposed to those which are *datum oriented* which go further in pinpointing the values likely to be wrong, and not only the points.

The problem is: given an event with w numbers, which one is wrong?. In other words, once a point in R^w space is selected, the value (or values) which make it unusual should be highlighted. 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 value is more likely to be in error. As a solution a sensitivity analysis has been sought using a functional S designed to highlight any unusual situation. S(k) is typically small if all the scores $a_i(k)$ are themselves not large. It is defined as

$$\mathrm{S}(k) = \sum_{\mathrm{i} \in \mathrm{p}} \frac{\mathrm{a}_{\mathrm{i}}^2(k)}{\mathrm{w}_{\mathrm{i}}} \tag{2}$$

$a_i(k)$ is the i-th score for the k-th outlying event and w_i is a weigth. The index i varies within a set p pointing to those scores considered as noisy. Hawkins (1974) used the associated eigenvalues instead of w_i, but we used the criteria suggested in *paper II* which make every term in de summation of the same order. In order to isolate the problematic station it is proposed to calculate for the event in question all the partial derivatives of the functional

$$\frac{\partial \mathrm{S}(k)}{\partial \mathrm{x}_{\mathrm{j}}} = \frac{\partial}{\partial \mathrm{x}_{\mathrm{j}}} \sum_{\mathrm{i} \in \mathrm{p}} \frac{\mathrm{a}_{\mathrm{i}}^2(k)}{\mathrm{w}_{\mathrm{i}}} \qquad \mathrm{j} = 1..\mathrm{n} \tag{3}$$

where x_j denotes the readings from the j-th station, p the set of noisy scores and $a_i(k)$ the i-th score for the k-th outlying event. The maximum derivative (in absolute value) will identify the most sensitive station, which will be taken as the error candidate.

The S statistics have been suggested for the first time by Hawkins (1974). It is a semi-distance, closely related to the Mahalanobis distance MD_i defined as

$$MD_k = \sqrt{(\mathbf{x}_k - \mathbf{T}(\mathbf{X}))\mathbf{C}(\mathbf{X})^{-1}(\mathbf{x}_k - \mathbf{T}(\mathbf{X}))^T} \tag{4}$$

being $\mathbf{x}_k$ the vector with the observed values for the k-th outlying event; $\mathbf{T}(\mathbf{X})$ estimated as the arithmetic mean of the data set $\mathbf{X}$ and $\mathbf{C}(\mathbf{X})$ estimated using the usual sample covariance matrix. If the set p includes all the scores, the statistic S(k) defined by Hawkins (1974) is equal to MD_k. The sensitivity calculations are carried out independently for each k-th outlying point in the R^w space. This completes the description of our method, first proposed in *paper I.*

4.1.4 Principal Component Analysis based methods (Hawkins 1974)

Hawkins (1974) directly used his statistics (denoted here as S) as an outlier detector tool, the larger values being associated with outliers. After some assumption regarding the multivariate pdf of the population, he is able to show that the Mahalanobis distance should follow a chi-squared pdf; using confidence levels, those events which produces MD_i values larger than a prescribed number (function of the dimension w and the confidence level) are considered outliers. Since the covariance matrix is positive definite, the surface described by MD_i=const. is an ellipsoid in R^w space, and all points inside it will not be considered outliers.

The non outlier region suggested by our method bounds within an interval only some of the scores; it can be represented as an hypercilinder, with a "rectangular" cross section. The open boundaries corresponds with the unrestricted scores. Any point lying ouside such a hypercylinder is regarded as an outlying point. Hawkins's criteria also led to a hypercylinder (because some scores not belonging to the set p are not limited), but with a second order cross section. Figure 4.2 shows the differences for the case w=3. The projection of both the Hawkins (1974) region and the Mahalanobis distance coincide in the plane (**e**$_2$, **e**$_3$) because only a$_1$ is unbounded. Notice that the axis are related with the scores (not the original values), and that we have shown only positive values for the a$_1$ score; also the relative size of the rectangular-like region has been changed only for the sake of the figure.

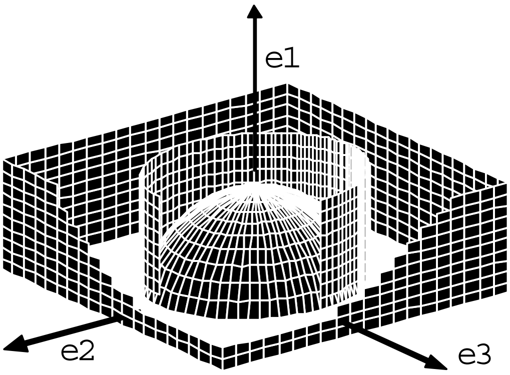

Figure 4.2 Sketch of the different regions for outlier detection used by the approach presented in paper I, the Hawkins74 one, and the standard Mahalanobis distance (from outside to inside) for w=3. Points are not considered outliers if they are in the inner part of the region. The score a_1 is not bounded for the two outer regions, while a_1, a_2 and a_3 are bounded for the Mahalanobis ellipsoid.

4.1.5 Artificial Neural Networks

All approaches for the error detection considered in this section in one way or another find a statistics which shows to be non-robust when calculated using outlying points. Large values are associated with unusual situations, which in turn are natural candidates to be outliers. Taking advantage of some aspects of the idea proposed for the scores we analyzed some results arising from the missing value problem (to be presented in 4.2) and in particular the Artificial Neural Networks techniques. ANN methods are rather new and thus we will give a brief outline here (see Warner and Misra 1996; Stern 1996 for a more thorough presentation).

ANN are based upon simple models of biological neural networks. There are different designs depending on the application. We have used it to fit a multivariate time series using available data, where both the inputs and the outputs are real numbers. In general, the ANN is organized in layers (see fig. 4.3), the first one being stimulated directly by the observed values. Each neuron of the next layer is stimulated by a linear combination of the outputs of the previous layers by means of a simple transfer function. For example, the logsig (Demuth and Beale 1994) function is given by:

$$out_j = \left\{ 1 + \exp\left[-\sum_i \left(a_{ij} * input_i \right) \right] \right\}^{-1} \tag{5}$$

with parameters a_{ij} (named synaptic connections) to be fixed for each neuron. The ANN requires, like its biological counterpart, a training process which is simulated here by means of adjusting the a_{ij} parameters. A bias term is also added to the weighted average of the inputs. Cybenko (1989) proves that, under some hypothesis, an ANN with finite number of neurons and one hidden layer can approximate any continuos function of n variables to an arbitrary extent, where n is the number of inputs.

We have designed and compared a number of architectures, which vary depending on the transfer function, the number of neurons in the hidden layer(s) and the input data. The terms *purelin, logsig and tansig* and its transfer functions are defined in Demuth and Beale (1994). Despite all of them might approximate a given function using enough neurons in the hidden layer, we want to keep this number low for practical reasons connected with training requirements. After some analysis of the problem, we realized that one important case to be covered is the identity function. In our problem, both input and output are homogeneous (i.e. share the same units) and at least on principle they might be the same if the measuring stations are close enough. So at least as a particular case, the ANN should be able to map easily the identity function. This cannot be achieved trivially with the available transfer functions in Matlab, because a maybe large number of hidden neurons is required. So new transfer functions *asinh* and *sinh* have been implemented in order to assure that the ANN can model easily the identity function, i.e. its output is equal to its input even with a single neuron in the hidden layer. According to Cybenko (1989) non linear functions could be also modelled as well.

Since the daily precipitation has a compound probability density function (pdf) with nearly 80 per cent of the readings equal 0, we attempted to transform it to obtain a more regular (nearly uniform) pdf, using its cumulative probability density function. It should be stressed that for each station a different ANN need to be trained, using the other stations as inputs. That imply 10 ANN in the daily precipitation case. For those schemas using data of a single day, there are 9 input values; where two days are used there are 10+9=19 values. The ANN named bp11 is a special case, because is used as a pure forecasting tool; its inputs are 10 values of the day before.

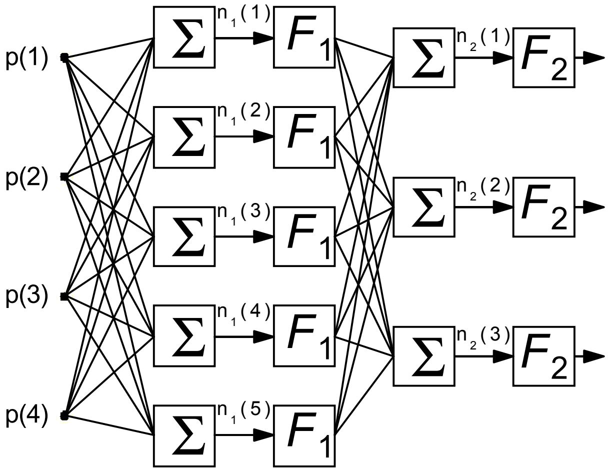

Figure 4.3 Sketch of a typical ANN organization. Information flows from left to right. There are four inputs p, one hidden layer with five neurons with transfer function F_1, a second hidden layer with three neurons with transfer function F_2 which produces three outputs. The summation symbol indicates a weighted average of all outputs from the previous layer plus a bias term $n_i(j)$.

In order to handle the possibility of more than one missing value for a particular day, arbitrary initial numbers are assumed. An iteration is performed in order to satisfy the output for all the ANN involved. All of the ANN were trained using one third of the available records (without missing values) trying to minimize the RMS of the error. This approach is named supervised learning (Warner and Misra 1996). The error is defined as the difference between ANN output and true value.

Training was done using backpropagation (Rumelhart *et al.* 1986) and for practical reasons, the number of iterations was kept below 200. The training cost in terms of CPU time is high: over 10 hours on SUN 20 for each meteorological station depending on the complexity of the ANN. This is not unique for this problem: Nychka and O'Connell (1996) state *"...In speeding up the fitting process we have found that intelligent strategies have failed miserably and brute computational force seems to carry the day. Currently, our preferred method is to use many (several thousands) random starting parameters and out of these pick the best hundred or so for a robust optimization with a fairly large tolerance for convergence. From this second set, we pick the best 20 for a high tolerance optimization.... In our experience this shotgun approach works well in providing high quality fits and anticipating the fact that the sum-of-squares surface has numerous local minima...".* We used almost the same strategy, limiting the random starts to a few dozens instead of the thousands mentioned. Increasing the allowable limit for the number of iterations showed little improvement.

The PCA method for error detection can be considered a particular case of ANN, using one hidden layer with linear transfer functions. The second layer has as many neurons as input variables, and each neuron is associated with one score. Its outputs are processed by neurons with binary output: if the i-th score is within a prescribed interval, the output is zero, implying no alarm. Otherwise, the output is one, indicating the possible existence of an outlier and the neuron is said to be *activated*. The net output of the ANN is obtained by a logical "OR" operator: if any of the third layer neuron is activated, the net output is one and outliers are believed to exist.

We have shown in *paper I* that extreme values for the noisy-related scores are associated with unlikely events; we detect those events by checking that at least one of those scores is outside of its outlier region. We applied the same idea to the general ANN with non-linear transfer functions. We attempted first to find the noisy neurons looking at the weights themselves. From the population we derive limits for the outlier region for the scores (i.e. outputs for the noisy neurons in the hidden layerh) and the method should flag those cases which do not behave like the overall population. The non linear characteristics of the transfer functions makes the dispersion of the population of each particular score very wide (even orders of magnitude), so unlikely events are clearly pointed out.

Assuming that we have 10 values for a particular day, we also have 10 different ANN which can predict any of those values using the remaining 9. Since we do not know in advance which is likely to be in error, we will use all 10 ANN and analyze the output of its hidden layer. Notice that each ANN has at least one hidden layer, whose weighted outputs are supposed to be added and used to predict the missing value. We assume that those neurons in the intermediate layer which have the highest weights are more related with noise, because typically the output is small. Thus we decide that a neuron in the intermediate layer explains mostly noise if its weight in the output layer is more than five times the smallest weight of that layer. After an analysis of the population we are able to find the outlier region for all those neurons and classify the event as unusual if at least one of the outputs falls outside the outlier region.

The "five times" limit has been determined by trial and error. As might be noticed, at present we are only able to classify the whole event as unusual; some more research are required in order to pick the value (or values) responsible for this behavior. This method is also *event oriented* (as defined in 4.1.3). Using the described criteria in order to identify the noisy neuron(s) we were able to test most of the available ANN (designed for an imputation task) for the quality control purpose. The exception are bp22 and bp23, because they are trained for each event separately and they will have a variable number of neurons (see 4.2.3 for further details).

4.1.6 Likelihood based method

The last new method to be described is based on Geostatistical concepts (Samper and Carrera 1990). The fundamental problem is to find a suitable interpolator given a finite number of observation points (which might belong to 2D or 3D space). Under some assumptions, in the classical kriging method (Samper and Carrera 1990) the interpolated value can be regarded as a weighted average of the available data, with weights depending on the coordinates. It is assumed that the data field is homogeneous (i.e. its statistical properties do not vary at least in the neighborhood) and they can be fully described by means of a variogram function. The isotropy (i.e. independence to the direction) is usually assumed or obtained via a transformation of the coordinates (Samper and Carrera 1990).

In order to find the variogram function, a number of methods have been proposed. Samper and Neumann (1989) proposed to find the unknown parameters of the variogram by maximizing the likelihood of the sample. In its simplest way, it implies that a) one data point is removed from the dataset b) it is interpolated using the other available data c) the difference between the original and the imputated value is stored. Once this procedure is performed for all or many of the data points, the likelihood of the set for the given parameters can be calculated and (according to Samper and Neumann 1989) it can be maximized under some general hypothesis.

In our case, we do not want to interpolate points others than the measuring net, and we follow the meteorological practice of Objective Interpolation (see section 4.2.2) in assuming that the statistical properties are valid for similar weather situations. Thus we can use all available data to calculate the sample covariance matrix, instead of obtain it via the variogram. As a practical consequence this implies that the variogram need not to be calculated nor modeled, and thus apparently we have no need for the likelihood.

Our proposed strategy is to calculate the likelihood for a given date using the covariance structure derived from the whole event population. We used the linear method coded as gandin20 (see section 4.2.2), which is standard in meteorology. The likelihood value itself is a measure of how likely or unlileky is the imputation obtained: since we are using the available data, one side effect is that low likelihood values might be connected with outliers on some of the records of the day. So we calculate the likelihood according to Samper and Carrera (1990), sort the records and suggest that those with the lowest values are the ones with errors. As before, the method is *event oriented*, and suggesting the most unlikely value within the event is still under research.

4.1.7 Mahalanobis distance-based methods

In addition to the new methods already described, we tested also a number of methods well known in the literature. For the sake of completeness a brief summary is included here. As mentioned before, the classical Mahalanobis distance is used as an indicator of outliers. It is defined for any set **X** and for any event $\mathbf{x}_i$ (Rousseeuw and Van Zomeren 1990) as

$$MD_i = \sqrt{(\mathbf{x}_i - \mathbf{T}(\mathbf{X}))\mathbf{C}(\mathbf{X})^{-1}(\mathbf{x}_i - \mathbf{T}(\mathbf{X}))^T} \tag{6}$$

being **T**(**X**) estimated as the arithmetic mean of the data set **X** and **C**(**X**) estimated using the usual sample covariance matrix. The distance MD_i tell us how far the $\mathbf{x}_i$ is from the center of the cloud. The covariance **C**(**X**) is a positive-definite matrix, so the set of events $\mathbf{x}_i$ with the same Mahalanobis distance lies on the surface of an ellipsoid with center **T**(**X**). Under some hypothesis large values for the Mahalanobis distance correspond to outliers; for normal distributions the squared Mahalanobis distance should follow a chi-square law.

However, calculating **C**(**X**) and **T**(**X**) with the standard procedure suffers from the *masking effect* which appears when a cluster of outliers is present. **C**(**X**) and **T**(**X**) are affected and the outliers no longer have a large MD_i. To overcome that problem, some other estimates of **C**(**X**) and **T**(**X**) have been proposed. The term "high breakdown" is coined in the statistics literature to express that the results will be unaffected even by arbitrary large errors in a fraction ε of the population. The theoretical bound for ε is dependent on the method, but in all cases it should be slightly less than half the population.

Among the high breakdown methods, we have considered the Minimum Covariance Determinant (MCD), the Minimum Volume Ellipsoid (MVE) and the Hadi's method (Hadi 1992, 1994) as well. All of them produce a robust estimation of **C**(**X**) and **T**(**X**). Once they are available, the Mahalanobis distance can be calculated for all events, and they can be ordered accordingly. Those events with larger distances will be the first candidates to hold outliers. Hadi (1994) suggested that under multivariate normal hypothesis, only those events with a Mahalanobis distance larger than a preset value should be considered as outliers. The preset value depends on the number of the columns and of a confidence level. In the simulations we ignored such limit and get new candidates from the ordered list as requested. In addition, since the estimators are robust, it will be useless to re-calculate them after removing some errors.

With this procedures we have a means to determine if an event is to be considered as outlying or not. However we have no suggestion about the individual records of the event, so we extended the application of the robust distance a bit further in order to have *datum oriented* methods. Using the robust distance, we suggest (as presented in *paper I*) to calculate the sensitivity of the distance to the data values, and use the most sensible values as a candidate for being an outlier. We made the experiments with both cases: the event oriented code check all values in a suspicious event while only those which are more sensitive are checked for the datum oriented case.

- **Minimum Covariance Determinant and Minimum Volume Ellipsoid**

MCD (Rousseeuw and Leroy 1987) searches for a subset of **X** containing nearly half of the data, the covariance of which has the smallest determinant. Since part of the data is ignored in computing this subset, the method can accommodate nearly 50% of the population with outlying values. Based on an idea exposed by Hawkins (1993) for regression applications Hawkins made a program for fast estimation of the MCD wich was used here.

The MVE (Rousseeuw and Van Zomeren 1990) algorithm find **T**(**X**) and **C**(**X**) in order than the MD_i for half of the population is below a prescribed constant which under multivariate normal hypothesis only depends on the number of columns of **X**. Since **C**(**X**) is positive definite, the MVE can be interpreted as an finding the center and the principal axis of an ellipsoid of minimum volume containing half of the points of the cloud (see original reference). However, **C**(**X**) is not the covariance of a subset of the data as in MCD but a weighted average of all the events. Both MCD and MVE are expensive in terms of CPU time: we were forced to limit the number of trials in our calculations, so our estimate of the true MVE and MCD might be improved.

In order to lower computer time, in the precipitation case we excluded from the **X** dataset those events with all-zero readings, and forced the ellipsoid to be centered in the origin. This will assure that all records with all-zero readings will be inside the ellipsoid, and thus we were able to use the program with substantially less events. For wind data, we simply subsampled the available dataset in two halves, and applied the method to one part only. The computations for MVE were carried out using the program MINVOL, available in statlib.

- **S and M-estimate for multivariate location and shape**

It can be shown that some of the previous methods are particular cases of more general estimators. Following Rocke (1996) we will define an S-estimate of multivariate location and shape as that vector **T**(**X**) and positive definite symmetric matrix **C**(**X**) which minimize det(**C**(**X**)) subject to

$$n^{-1}\sum \rho\left(\sqrt{(\mathbf{x}_i-\mathbf{T}(\mathbf{X}))\mathbf{C}(\mathbf{X})^{-1}(\mathbf{x}_i-\mathbf{T}(\mathbf{X}))^T}\right)=b_0 \qquad (7)$$

where ρ is a nondecreasing function on [0,∞]. The function ρ is usually differentiable (the major exception being the MVE, where ρ is 0 or 1). For the case of the multivariate normal distribution $\rho(x)=0.5x^2$, and Rocke (1996) states that this function should not vary with the number of variables.

The M-estimate can be defined (Maronna 1976) as a vector **T**(**X**) and positive definite symmetric matrix **C**(**X**) which is a solution of

$$\sum(\mathbf{x}_i-\mathbf{T}(\mathbf{X}))u_1(d_i)=0 \qquad (8)$$

$$n^{-1}\sum(\mathbf{x}_i-\mathbf{T}(\mathbf{X}))(\mathbf{x}_i-\mathbf{T}(\mathbf{X}))^T u_2(d_i^2)=\mathbf{C}(\mathbf{X}) \qquad (9)$$

being u_1 and u_2 non negative, non decreasing functions for positive arguments, and d_i defined as

$$d_i^2=(\mathbf{x}_i-\mathbf{T}(\mathbf{X}))\mathbf{C}(\mathbf{X})^{-1}(\mathbf{x}_i-\mathbf{T}(\mathbf{X}))^T \qquad (10)$$

The high breakdown properties of both estimators have been analyzed in a number of papers (see Rocke 1996 for a review). Rocke and Woodruff (1996) have implemented a code available in statlib which uses some particular functions u_1 and u_2 after analyzing choices from Rocke (1996).

- **Hadi's method**

The method by Hadi (1992, 1994) produces a result formally similar to that of the MCD. It attempts to find a subset containing nearly half of the population with the minimum distance to an estimation of the covariance matrix. However it is not combinatorial, and evolves from an initial estimate by adding a new point until the required set is obtained. The algorithm is considerably faster than the others, but it suffer from some drawbacks: it is not affine equivariant (Rocke and Woodruff

1996) which implies that the results are different after a linear transformation of the data. Hawkins' hypercilinder, MVE and MCD are affine equivariant.

4.1.8 Combined Mahalanobis distance-based and regression methods

Another possibility was described by Rousseeuw (1991) in the context of regression. He suggests that outliers can be detected by means of their discrepancy from a robust fit. This procedure is *datum oriented*, and can be considered a generalization of the rule valid for 1-D data with normal distribution: look for those values x_i which make $|x_i\text{-mean}(x)|/\sigma$ larger than a preset threshold (mean(x) stands for the sample average, and σ for its standard deviation). In the multivariate case, $1/\sigma$ is substituted by the inverse of the sample covariance matrix, leading to the square root of the Mahalanobis distance. This procedure suffers from the *masking effect* if more than a single outlier is present. Rousseeuw suggests to use robust estimates of the covariance matrix, like the MCD or the MVE defined above.

Once a robust estimate is available, we are able to perform a regression for event i of $\mathbf{X}_i(j)$ in terms of $\mathbf{X}_i(k)$, $k \neq j$, $i=1..n$. Let us consider a fixed column j. The robust residual can be computed as the difference of the $\mathbf{X}_i(j)$ and its robust estimate, divided by some estimator of the residual scale. *Independent of the regression model* one can calculate the robust distance to the center of the ellipsoid defined by MVE or MCD, which gives an *event oriented* index for outlyingness. If we define (following Rousseeuw) a *regression outlier* as the event which violates the regression model suggested by the majority of the data ($\mathbf{X}_i(j)$ is obviously considered) and a *leverage point* as a case i in which $\mathbf{X}_i(k)$ is an outlier with respect to the bulk of the data ($\mathbf{X}_i(j)$ is not considered) we will have a clear indication wether or not $\mathbf{X}_i(j)$ is an outlier (provided that it is the only one in the event!). A leverage point in 1-D is one which has a x-value very different from the population; case (c) in figure 4.4 shows that it is not necessarily also a regression outlier.

We described this method for the sake of completeness, despite it has not been considered in our experiment because it requires to analyze visually all Monte Carlo simulations. However, it is believed that it will be certainly of interest for GIS users dealing with a single sample.

4.2 The Missing Value Problem in tabular quantitative data

4.2.1 Aim

As mentioned before all mathematical models used in GIS environments rely on the availability of *suitable* data i.e. the data should be relevant, accurate and complete. With the term complete we mean that data has been sampled to an extent that makes them useful. For some cases, one might re-sample (interpolate) the available data to a prescribed grid in order to feed the models with the appropriate field, despite it is possible (at least on principle) to go to the ground and take a direct measurement. The interpolation of evenly distributed data with spatial coordinates is the topic of geostatistics.

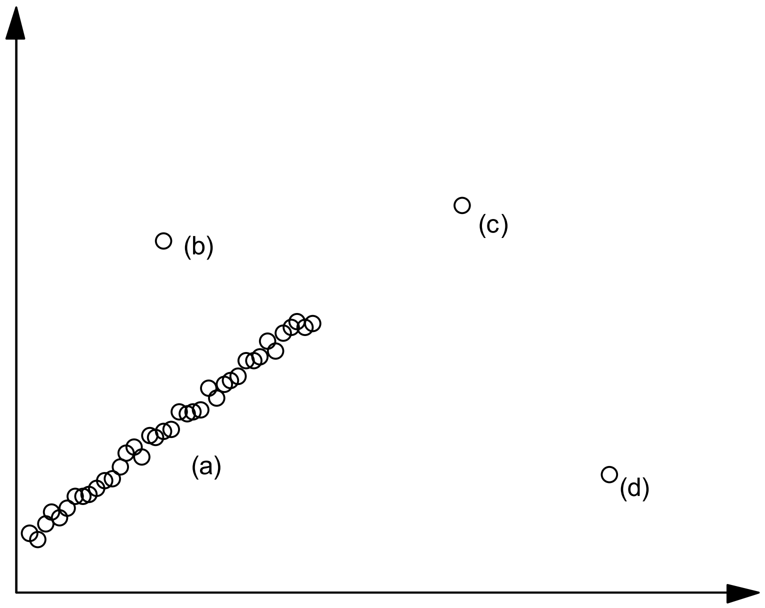

Figure 4.4 Simple Regression Example with (a) regular observations, (b) vertical outlier, (c) good leverage point, and (d) bad leverage point. Taken from Rousseeuw and Van Zomeren (1990)

We consider the case of missing records in meteorological data, which are sampled at carefully defined intervals in time on a limited number of locations. Such time series have the undesirable property that if a measurement has not been taken, it cannot be taken anymore. Depending on the mathematical model, it might be required that the missing value should be substituted with an appropriate estimate. The older *paper II* describes early developments on the subject, while *paper III* can be regarded as a methodological continuation of the former.

The different methods can be divided into two categories: *linear* or *nonlinear*. In the first case the estimated quantity is a linear combination of the available data. Its general expression is

$$y_j=\mathbf{wx}+b \tag{11}$$

being y_j the unknown quantity, **x** a vector which entries are the available data and b a scalar constant; both the weight vector **w** and the number b depends on the method. Typically the vector **x** holds the values of the same day, and both **w** and b are constants for the whole dataset at least for a period; for example, they can take different values for different seasons. In the second case the value is assigned using a non-linear formula. They proved to be the most interesting ones, despite their heavier demand on CPU resources. We compared over 40 different methods, and due to space constraints we will summarize those who proved succesful.

4.2.2 Linear methods

Due to their simplicity, these methods are widely used. This category includes among others, the methods of Cressman, Optimum interpolation (also named kriging), Ordinary least squares, as well as other simpler ones, like the nearest neighbor.

- **Cressman**

The requested number is obtained after a linear combination with weights which are the inverse of squared distance. The method does not require historical information, but only the station coordinates. The number b is zero and the entries of vector **w** are calculated as

$$w_i = \frac{1/d_{ij}^2}{\sum_{k\in N} 1/d_{kj}^2} \tag{12}$$

- **Optimum interpolation (Gandin 1965; Johnson 1982)**

This method is routinely applied for the specification of initial conditions for use in global weather forecasting programs. Instead of interpolate the desired field, it interpolates the anomaly or difference with a simple predictor, and the spatial correlation properties of the anomaly field are analyzed. Usually the anomaly is assumed isotropic and homogeneous, and it should be modelled at arbitrary (x,y) in the general case of interpolation. However, if the point where the prediction is required belongs to the measured point set, its covariance with the other stations is available. For this particular case this method (also known as kriging) is equivalent to the Ordinary Least Squares. The covariance might be calculated separately for winter and summer, or for all together as we did. This procedure allows information from the day before to be used. The procedure can also consider information on known accuracy of the dataset, or estimate it as well as described in Johnson (1982).

For the tested datasets we used different anomaly fields and transformations of the original values which are summarized in table 4.1. For example, the method coded as “gandin7” assigns values for the variable *x_i=sqrt(precipitation$_i$),* computing the anomaly with respect to the historical mean. In this case $b=mean(x_i)$ (mean stands for average over time). The classic Optimum Interpolation procedure is coded as "gandin20". Methods denoted as gandin6 and gandin4 differs in the simple predictor used for their anomaly.

Because daily precipitation has a very irregular probability density function (pdf) we designed a transformation $x_i=f(precipitation_i)$ which makes pdf(x_i) nearly uniform, except for $precipitation_i=0$. The transformation based on the cumulated density function has an inverse and assures that x belong to the interval [0,1]. Another detail that should be commented is regarding the handling of missing values. We noticed that around 30 per cent of the events have at least one missing value. So, in principle, they cannot be included in the calculations of the covariance matrix. We decided to use as much information as possible and calculate the covariance matrix in an iterative way. For each of the methods defined in table 4.1, we first estimate the covariance matrix from the full sampled events. We then apply the corresponding gandinX method for imputation, complete the dataset as much as possible, and update the covariance matrix using all the events now complete. We continue until there is negligible change in the imputed values. We have considered in early stages to calculate the covariance matrix componentwise, using all events with data in station i and station j to calculate the entry (i,j). This procedure led to a non-positive definite covariance matrix, and the approach was discarded.

Table 4.1 Brief information about the different methods based on climatological functions. f(precipitation) denotes the transformation which renders a nearly uniform probability density function(see text). t and t-dt denotes values from the day and the day before

Our coded name	*Anomaly*	*Variable*	*Using data from days*	
	respect to:	*to interpolate*	*t*	*t-dt*
gandin	*historical mean*	*precipitation*	*X*	-
gandintrans	*historical mean*	*f(precipitation)*	*X*	-
gandin6	*historical mean*	*precipitation*	*X*	*X*
gandin7	*historical mean*	*sqrt(precipitation)*	*X*	-
Initial value for the field chosen as zero				
gandin_diario	*0*	*precipitation-daily mean*	*X*	*X*
gandin4	*0*	*precipitation*	*X*	*X*
gandin5	*0*	*precipitation*	*X*	-
Neglecting instrument error				
gandin20	*historical mean*	*precipitation*	*X*	-
gandin3a	*historical mean*	*precipitation-daily mean*	*X*	-

- **Ordinary Least Squares (OLS)**

This is a standard method and the theory for it can be found elsewhere (Dahlquist and Bjork 1974). The weights **w** are chosen in order to minimize the 2-norm of the vector $\mathbf{M}^{(j)}\mathbf{w}$-$\mathbf{m}^{(j)}$ (a scalar proportional to the Root Mean Square of the Errors, RMSE) where $\mathbf{M}^{(j)}$ is the matrix of the available data (as many rows as dates, as many columns as stations but without the j-th one) and $\mathbf{m}^{(j)}$ is a column vector with the j-th stations values. The version implemented assumes that the data is error free, so **w** can be derived from (dropping the index j) $\mathbf{M}^T\mathbf{M}\mathbf{w}=\mathbf{M}^T\mathbf{m}$, b=0. Notice that this method is prone to suffer from the existence of outliers; the remedy is either to remove the outliers before the calculations or to use an estimate more robust like the ones described below. It should be mentioned that for this and the following methods we used events with no missing values only, as opposed to the procedure used for the gandin-like methods. The method denoted as "gandin20" is equivalent to OLS if there are no missing values.

- **Least average (Minimum 1-norm)**

Here the weights **w** are chosen in order to minimize the 1-norm (sum of absolute values) of the elements of the vector $\mathbf{M}^{(j)}\mathbf{w}\text{-}\mathbf{m}^{(j)}$, a problem involving the solution of a non-linear optimization task. For OLS and gandinX it was only required to solve a linear sistem of equations, so calculate the weights **w** consumes substantially more CPU time than all previous methods. However, it is more robust against outliers.

- **Least 95 percentile**

Since the population might be affected by a small set of gross errors, it is fit to minimize a robust statistic, as the 95 percentile of the distribution of absolute errors. As before, calculation of weights **w** for this method consumes significant CPU time.

- **Least Median of Squares (LMS)**

Rousseeuw (1984) suggested for the first time to use a robust (i.e. outlier resistant) estimator for regression instead of the sum of squares of the residuals. He suggested to use the median of the squares of the residuals, and a number of widely available FORTRAN programs have been developed to calculate it. Such programs are well suited for small datasets, and the computing time required to obtain the coefficients in our case is considerable. Hawkins (1993) described an algorithm to perform the calculations, who also implemented and kindly furnished the program. The results were 10 different sets of coefficients, assuming that only one station is missing. For the case of more than one missing value in a particular event, we proceed in an iterative fashion as sketched in figure 4.5. Any finishing criteria can be used; we requested that the maximum discrepancy between estimated and calculated missing values is below a threshold.

```
assume an initial value for all missing values
repeat until finishing criteria is satisfied
     foreach missing value
          calculate the regression using most
          recent estimations
     end foreach
end repeat
```

Figure 4.5 Pseudo code for the imputation of more than one missing value

The method is said to have a *high breakdown,* because its results are insensitive to arbitrary large outliers, provided they are no more than nearly half of the population (Rousseeuw and Van Zomeren 1990). To decrease CPU time, we assumed that there are not more than 10 per cent of outliers in the population. For the precipitation case we extracted from the dataset those days with all-zero precipitation readings, and we required that the regression line should go through the origin (in order to easily accomodate back the case of all-zero readings).

- **Least Trimmed Squares (LTS)**

This criterion was suggested for the first time by Rousseeuw (1984). The weights **w** are chosen in order to minimize a weighted sum of the squares of the residuals, the weights being either 1.0 or 0.0. There are as many zero weights as the number of outliers assumed. The method has high breakdown, and its results can be extended up to a breakdown level similar to that of LMS. Hawkins (1994a) suggested an algorithm for calculate the coefficients, implemented and kindly supplied it. Once calculated, weights **w** were used with similar criteria as of LMS.

- **Nearest Neighbor**

In all cases the missing value is taken directly from another station following a given order. All weights are zero, except one which is 1, and the number b is 0. We considered three criteria for the distance. In the first case, the order is based on euclidean geometrical distance, and in the second case we used the expertise from a meteorologist which analyzed qualitative similarity. The third alternative was due to Gutiérrez (1996) who selected the nearest station taking into consideration the Kulback-Leibler's (Borovkov, 1987) pseudo distance between the station's probability density functions.

- **Assign a constant value**

This is a simple method, which disregards any other information. We applied it using the modal value and the expected value, and to get an idea about the characteristics of this dataset.

- **Assign the daily average**

This is almost self explanatory; the entries of vector **w** are all equal and the number b is 0.

- **Univariate time series interpolation**

This is simple method uses only data from the station under consideration, disregarding any correlation with the others. In our implementation, we used the nearest (in time) available record before and after the missing value, and a linear interpolation in time is used to predict the missing value.

- **Temporal Interpolation of Principal Scores (TIPS)**

This method were proposed in *paper II* and assumes that only those scores related with the structure of the cloud contain information relevant for recovering it. After an analysis of its time series properties we realize that those scores has most of its energy in the low frequency range, while the noisy scores span it through all the spectra. This fact was observed both for the daily rain and the hourly surface wind records and is believed that is a general rule. The method will interpolate the time series of the scores related to structure, and set as zero those scores related with noise. Once the scores are calculated this way, all the values for such event can be calculated. The missing values can be recovered from them, and using the observed values the scores can be recalculated for the same day using now all the available information. If the time gap is larger than one unit, the procedure can be used as described in *paper II*. Notice that this is a multivariate time interpolation procedure which will be equivalent to the standard univariate case if all the scores are classified as structural.

- **Penalty of the Principal Scores (POPS)**

This method were also proposed in *paper II.* If we analyze the histogram of the scores a_i it can be observed that the main PC has a pdf which is heavily skewed or has a significant dispersion around zero. On the other hand, for the weak PC the histogram is symmetric and the dispersion around zero is very low. Any imputation procedure should preserve this properties, and then it should produce scores a_i consistent with this histograms, i.e. very near zero for all weak PC. Such property might be imposed as a condition, choosing for any given date t all missing components of vector **P**(t) in order to minimize some penalty function, like

$$S(\mathbf{P}) = \sum_{i=k}^{i=n} w_i . a_i^2(\mathbf{P}) \tag{13}$$

being the scores $a_i(\boldsymbol{P})$ corresponding to the vector **P** (now complete) and the weights w_i selected in order to consider the different absolute value of each score a_j. The sum is taken over those scores previously classified as weak (or noisy) as described before. Vector **P** is only partially known, and it is assumed that it has q unknowns (or missing values). The optimum of S(**P**) can be obtained making its partial derivatives null for all unknowns

$$\frac{\partial S}{\partial p_{m(j)}} = 0, \ \ \mathrm{j} = 1..\mathrm{q} \tag{14}$$

being $p_{m(j)}$ the missing records for event t. The so defined linear system can be easily solved by standard procedures.

4.2.3 Non linear methods

- **Hotdeck**

The hotdeck procedure is typically used for surveys, and the missing value is assigned from other survey with the same (or nearly the same) answers in all the other available fields. The missing answer is taken from that survey which answers more closely matches the one considered. If there is no other survey which exactly matches, a distance between surveys has to be defined. If there are more than one survey which match, a random selection is performed. We simply used the sum of the squares as distance.

- **Artificial Neural Networks (ANN)**

In addition to the already described ANN architectures (see 4.1.3), a different solution inspired by the work of Kanevsky *et al.* (1996) was implemented as ANN bp22 and bp23; instead of using fixed values for the weights and bias in the ANN for all events we attempted to train the ANN in order to provide full interpolation capabilities (i.e., an output for arbitrary geographical coordinates) with a different interpolator for each day. In these cases, the input to the net is the x,y coordinate of the available stations (provided they are not too few) and the training process attempts to fit the available readings. After training, the ANN is supposed to approximate the function *precipitation(x,y)* well, thus it is asked to predict a value for those coordinates with missing values. We did not attempt to modify the basic routine available in MATLAB for training, which use a variant of steepest descend, and minimizes the RMS of the errors.

4.3 Quality control of raster datasets

4.3.1 Aim

Raster datasets are very common in GIS applications. Despite the fact that remote sensing data fall into this category, we decided to analyze the more classical case of DEM, mainly for the (at least theoretical) possibility to improve its accuracy to an arbitrary extent. However, the work of Maronna and Yohai (1995) should be quoted here, who analyzed 38 pixels on five frequency bands from satellite measurements in order to find outliers. They noticed that standard methods of regression (i.e. neglecting outliers) applied to this dataset will fail in discriminate between two clusters present there. Each cluster should have a different regression model, and blind application of the algorithm led to wrong conclusions. Maronna and Yohai applied two robust methods for outliers detection (one of them MVE) and showed their ability to separate such clusters.

After a review of the literature we found that blunder location methods typically are designed to detect errors in the production stage. The situation of the data producer and/or the end user attempting to locate the worst errors in order to improve the quality of the dataset at later stages is not considered. The aim of the study was to develop and test a method suitable for outliers detection in the final product. We omit any discussion about the genetics of the errors (i.e. its relation with the particular production procedure), because the end user might not be aware of the lineage details. The terms outlier, gross error and blunder are used synonymously throughout this section.

4.3.2 The method for error detection

In *paper IV* we designed and tested a method using a Monte Carlo procedure. We seed an available DEM of unknown quality with synthetic errors of low spatial correlation, and obtained good but preliminary results. The method has some parameters free, and after the simulation we were able to provide some rules of thumb in order to proceed in a different case. However the validity of the rules of thumb still depends on the error model selected. From the beginning we attempted to use or adapt those methods to the raster case which have proven to be sucessful with tabular datasets. Clearly, some tricks are required to do so and this section is devoted to show how it has been done. Clearly once we were able to apply one method, we could apply all of their equivalents.

Even though our procedure is based upon Principal Component Analysis, our approach is very different from the one typical in image processing. We devised a method for selecting unlikely profiles for a narrow DEM (i.e., when its length is substantially greater than its width) and we were also able to detect the best candidate within each profile. We then considered any DEM as composed of a number of narrow DEM (also denoted as strips; see fig. 4.6). Each strip is assumed to have length n and width w (w<<n). The method considers the strip as points in the R^w space.

We denote as profile a section of length n, and cross-section as the sections of size w. The case of w=3 were illustrated in figure 4.1 where each point M_k represents a cross-section. Since the stripping process can be done row-wise as well as column-wise, we construct the candidate set of outliers from the intersection of both. The strip can be directly associated with a tabular dataset, and any of the methods already presented in section 4.1 can be used. The procedure can be applied iteratively, since, once an error is detected and "corrected" the cloud is modified to some extent, and so the scores. We keep track of the points already checked in order to avoid to select them twice.

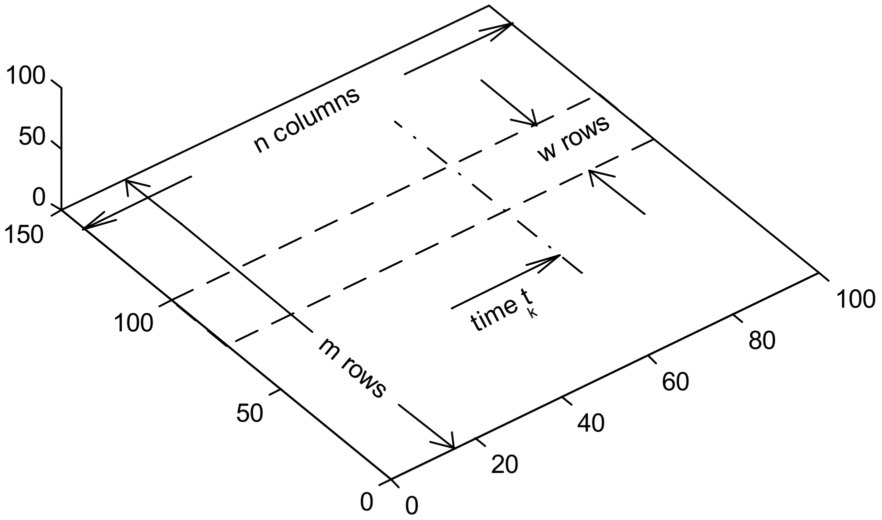

Figure 4.6 Sketch of the strip notation

The error location procedure directly analyzes the cloud of points in R^w, disregarding any order among points. This is an important assumption, since the concept of spatial self correlation completely looses all significance in the cloud. Adjacent profiles need not to be in any special order, since they are coordinate axes in the space R^w. The use of the cloud is in line with the common practice in statistics (Hadi 1992, 1994; Hawkins 1974, 1994a, 1994b, etc.) since the notion of "spatial correlation" and "precedence" is meaningless in most tabular data. The procedure involves five actions, and has been outlined in fig. 4.7. In the pseudo code we have used a single strip width w for rows and columns. We also assume that w is a divisor of m as well as n; this is not required.

In order to be able to perform a Monte Carlo simulation, some model for the error is required in order to produce realizations of the DEM with error. For the case considered in *paper IV* errors within [-4 m,+4 m] are typical. Since the data is rounded to the nearest meter, there will be little chance to pick errors of one meter.

Any "feasible" error should also be an integer number. The shape of typical (real) errors (i.e. its spatial structure) have received little attention in the literature, because common practice reports accuracy as RMSE, mean of the absolute value or percentiles of the elevation errors. Accuracy in terms of slope might be of help, but they are barely reported (Giles and Franklin 1996). We follow some authors (Bethel and Mikhail 1984) in modeling errors as additive and isolated, being the error elevation chosen from a given set. As a feasible set we have used, as a first example, the values [-4,-3,-2,-1,+1,+2,+3,+4] meters, with equal probability, which is considered a difficult case. This is expected to model spatially uncorrelated errors, and we denote it as spike-like errors (see fig. 4.8, left).

```
Given a DEM as a matrix of size m*n, subdivide the DEM
in row-wise and column-wise strips of width w
repeat until criteria is satisfied:
     a) increment previous  row-wise candidate set:
          a.1.- locate the columns likely to have
               candidates
          a.2.- within each column, find the rows that
               identify the candidates
     b) increment previous column-wise candidate set:
          b.1.-locate the rows likely to have
               candidates
          b.2.- within each row, find the columns that
               identify the candidates
     c) intersect both sets
     d) evaluate criteria
     e) correct all errors
end
```

Figure 4.7 Sketch of the steps required for the location of outliers in a DEM (from paper IV)

We seeded the DEM with this synthetic errors changing 5 per cent of the points; this value looks somewhat high when compared to the one reported by Östman (1987). He found a typical value of 0.5 per cent for the number of gross error occurrences, but here the worst errors are of absolute size 4m and they account for only 1/4 of the errors. As another alternative for an error shape model, we also tried a more structured one, which resembles a pyramid; once a point is selected, it is modified by adding a 2Δ error, and only Δ to the eight points surrounding it (see fig. 4.8, right). We have selected Δ uniformly from the set [-2,-1,+1,+2]. We named this model pyramid-like, and it is expected to model some degree of spatial correlation in errors.

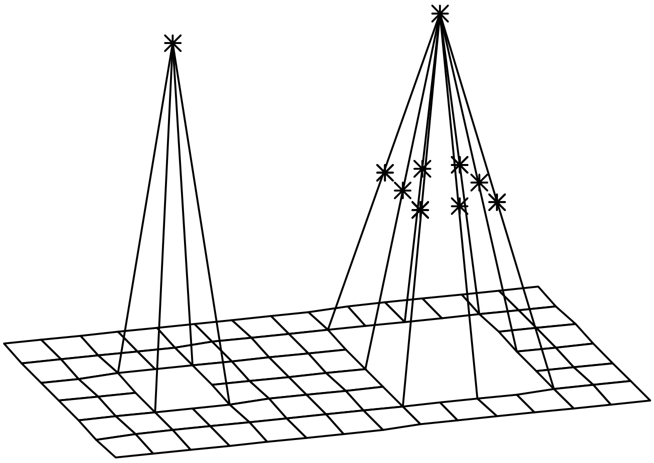

Figure 4.8 Sketch of the spike-like and the pyramid-like error model. An asterisk indicate modified elevation values (from paper IV)

4.3.3 The modified error detection method

The generation procedure for DEM usually assures that errors are correlated in space (see for example Day and Muller 1988). Thus we modified the proposed procedure (described in *paper IV*) somewhat in order to handle the correlation in space. Notice that the procedure of *paper IV* has been tested with synthetic, weakly correlated errors. It will be shown that its performance decays as the correlation increases. The procedure of Felicísimo (described below) suffers from the same problem, since the error at i,j is highly correlated with the one at the immediate neighbors. The method of *paper IV* does not require that the *along the strip* profiles are contiguous. Therefore we can skip some of them (the ones most correlated) for the analysis. The strip is chosen as before, but in the calculations we consider subsets created using every k-th row, k being related with the range, a geostatistical property (see Samper and Carrera 1990) of the error field. In *paper V* we assumed that the value of the range can be estimated from an independent analysis: it might depend on the DEM characteristics, method for obtaining it, scale of aerial photography, etc.

4.3.4 The method of Felicísimo (1994)

This method is the simplest one available for this problem. Assuming that outliers are only locally correlated, the method analyzes the differences $\delta_{i,j}$ between the elevation value $z_{i,j}$ and an interpolated guess $Z_{i,j}$ obtained from its immediate neighbors. Assuming that the difference $\delta_{i,j}$ has a Gaussian distribution with mean Δ and standard deviation S_δ (both obtained from the sample) a Student t test can be applied to validate the hypothesis that $\delta_{i,j}$ belongs to the population of deviations. Operationally, we analyze the statistics $t_{i,j}=(\delta_{i,j} -\Delta)/ S_\delta$ which can be interpreted as a standardized deviation.

We used a best fit approximation with a biquadratic polynomial using the eight closest neighbors to calculate $Z_{i,j}$. Along the borders we assume mirror symmetry, and in the corners we used a linear interpolation with the three closest values available. We point out as a candidate error value any $\delta_{i,j}$ that makes $t_{i,j}>3.219$. Felicísimo states that even though a significantly high value of $t_{i,j}$ does not necessarily imply an error, it is an excellent alarm sign. The method is very simple and is also parameter free.

Once an error is located and corrected, both statistics Δ and S_δ change and new candidates appear. The method can be iterated and it might stop if no more "outlying" values remains. This is undesirable because we know that there still are errors are in the dataset, so we proceed by lowering the significance level at least once. The new candidates, once corrected, modify the statistics, and new candidates with the previous significance level appear. We stop the iteration when a prescribed effort has been achieved.

4.4 Tabular qualitative data: the national census example

4.4.1 Aim

The aim of the study presented in *paper VI* was to develop a nearly on-line quality control algorithm suitable to be used in connection with the scanning procedure task after the 1996 National Census of Population and Housing of Uruguay (population ~3 million, houses ~200.000). Since scanning of such large amounts of paper (10^6 sheets) was attempted for the first time in the world for a national census, automatic quality control procedures were sought to minimize the typist correction effort while keeping the number of errors below prescribed values. The proposed procedure was designed for categorical answers only. Separate routines were used for the Optical Character Recognition task, where both numerical (quantitative) answers and handwritten text were expected. Since the automatic recognition of marks gave good enough error levels, the proposed method were not used in practice, but the prototype has been tested through a Monte Carlo procedure using synthetic errors.

4.4.2 The rule based method

There is little guidance in the statistical literature for this problem. The best known method is due to Fellegi and Holt (1976). They presented a method specifically suitable for qualitative or categorical data. It is based upon the existence of rules which relate the different fields in each record. Such rules should be formulated by experts, and express their judgment that certain combinations of values or code values in different fields are unacceptable. If a particular record does not satisfy one or several of those rules, the field (or fields) that contribute to them are rejected. Notice that this procedure relies on the existence of explicit rules (and experts behind them).

4.4.3 The Duplicate Performance Method (DPM)

Following the main idea of this thesis, we focused our research on methods which are independent of the data description itself. The above mentioned rule-based method is not suitable, because we need to build new rules for each new dataset. A general procedure for locating errors in a typing process is the Duplicate Performance Method. If data are typed (keyed in) twice and independently, and if the results are compared by a method that can be assumed error free (such as a computer program comparing files after data entry) and if all the disagreements are corrected, then the only errors remaining in the data set are those where both staff members were in error. If the ratio of disagreements to total items is low, then the individual error rates of both persons are low, and the probability of joint errors (the product of the probabilities of individual errors) is even lower (Strayhorn 1990). The method is extremely simple, and can be applied to any kind of data, both quantitative or categorical. Despite its simplicity, it has some desired properties:

- The probability of locating an error is independent of the error itself, so gross errors will be corrected as well as subtle ones. This will help in keeping the statistical properties of the database.
- It is also independent of the *order* the retyping is performed, so if only a fraction of the dataset is retyped, typically the same fraction of the errors will be corrected.
- The procedure does not require a large database, so it can be applied also to small ones.

For the sake of simplicity, we will assume that by typing a record twice all errors are removed ("perfect inspector" hypothesis). This will help us in simplifying some arguments, and the reader will easily notice that this is not a key hypothesis.

4.4.4 The method proposed in *paper VI.*

A new method which is able to improve the standard DPM by sorting the records putting first the most unlikely ones have been presented in *paper VI.* We will outline here how we handle tabular qualitative data in order to use tabular quantitative methods. As before we will use our workhorse based on PCA (described in 4.1.3) as an example. We will consider only the problem of selecting a specific event (a single survey) on the basis that there is something in the answers that make it unusual. Such an event should be retyped. In a real processing environment, if the record is still unusual, it will be carefully analyzed by a trained specialist, which may found (or not) reasons to reject or modify some answers in the particular survey. The method can be used to give the specialist a smaller selected set, with higher probability of having errors. Notice that this procedure will decrease the variability in the data, because "reasonable" errors values are prone to be ignored. It will be also assumed that all the answers and their alternatives have the same relative importance.

In categorical data, the codification procedure usually generates a set of valid values for each question. Those values are usually coded as integers, but the value itself is meaningless. In order to manage categorical data with our proposed method, we should translate such integers in a way that the results do not depend upon changing the order of the alternatives in the question or changing the codes.

The definition of outlier in categorical data may differ from that for real-valued data. There are no possibility of arbitrary large errors. It is also assumed here that the dataset has passed successfully some trivial logical tests, which pointed out for example, more than one mark in mutually exclusive answers, or similar things. Also all the coded values are within their prescribed ranges. These logical tests are very crude, and certainly should not be confused with the edits designed by experts in the particular data (Fellegi and Holt 1976). It should be regarded more as a computer specification for the data, rather than a quality control procedure.

Given the data, the corresponding question list and the feasible options, the user should eliminate those fields which are *a priori* uncorrelated with the others. Typical examples for survey data are all the information related with the zip code, city code, address, etc. Also numerical quantitative data should not be considered (for example: age, area of the building, etc.) except if a categorization is applied.

The dataset is usually available in table format, one individual per row, and one question per column. In order to have a numerically useful representation, we will binarize the dataset, creating a new table containing only 1 or 0. This also make the data homogeneous (dimensionless). In order to binarize the dataset, one may think on a multiple choice sheet. For any particular question, there are room to choose between some (maybe mutually exclusive) alternatives. Instead of coding a single number for the answer, we may equally store all the alternatives, putting a 1 or 0 if the option is true or not. In other terms, each column of the original table expands to as many columns as alternatives in the question, allowing only 0 or 1 as an answer. After repeated for all questions, the data (without missing values!) is binarized and presented in table (or matrix) format, any method suitable for quality control of quantitative tabular datasets might be used. We applied our workhorse based on PCA. Notice that the dimension "n" of the covariance matrix is not the number of controlled questions but the sum of all the options within them. It is assumed that the population is large enough to represent properly the true covariance with the sample´s covariance matrix.

It should be pointed out that, even in numerical datasets, usually the mean value and the Principal Components are real vector values, and so are the projections of the dataset on the PC, which here are called scores. That holds true even if the data are integer or binary numbers. For example, in a precipitation dataset, all values are integer and positive, but the scores are real, i.e., they belong to a different number category. When considering categorical binary answers a similar situation arises. Even though they are real, the possible values of the scores are limited due to a combinatorial constraint. We are implicitly requiring that this finite number is a big number (in the experiments, 2^{69}-1) because if the number "n" is low the distributions will not look like those of continuous data. Notice that the real valued scores are not arbitrary because they arise from a finite number of possible answers.

Since the matrix is range-defective due to the logical interrelationships between mutually exclusive answers, there will be some zero eigenvalues which imply that some scores should be (at least theoretically) identically zero. This makes a slight difference as compared with the situation for quantitative data (Hawkins 1974; *paper I*) where the covariance matrix is strictly positive definite. Once the sample distribution of the scores is created, confidence limits can be calculated. These values will define the outlier region (Davies and Gather 1993) but without assuming any particular pdf shape. Why do we claim that this is the outlier region?. Fig. 4.9 shows the sampled probability distribution function for the given database of some of the scores and the arrows point to two values: those marked with an "o" correspond to the original answers for a particular survey; those marked with an "x" are related to the same survey, but now contaminated by modifying one of the answers. In this particular case, it was imposed that the house is equipped both with a color and a black and white TV set, while originally it has only black and white. Notice that the effect is important mostly in the "weakest" scores (i.e. those associated with the lower non-zero eigenvalues of the covariance matrix) and that the ones associated with the "strongest" ones are only minimally modified. The proper limit between the "weakest" and the "strongest" is to be determined, and some guidance is given in *paper VI*.

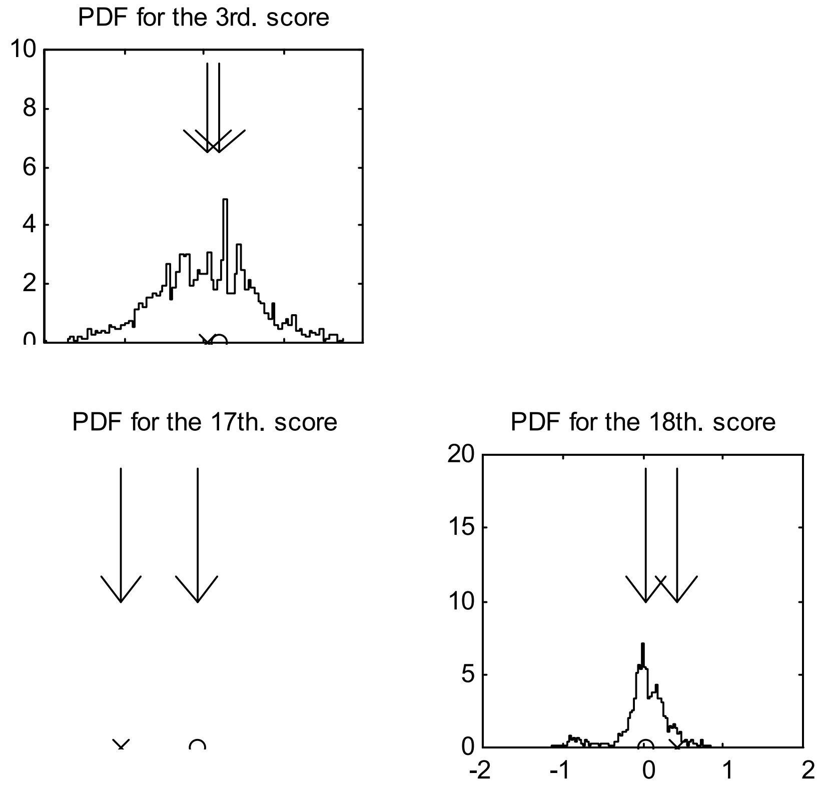

Figure 4.9 Example of the effect of a single outlier in a particular survey

5 Experimental setup and results

In this chapter we describe the experimental setup and results of the individual studies, as well as recent material.

5.1 Quality control of quantitative tabular data: the daily precipitation example

This example has been considered for the first time in *paper I* which has some methodological shortcomings. Further research has not been published before, and the results will be presented here.

5.1.1 Experimental strategy and measures of success

In *paper I* we applied the a PCA-based method to a single set of simulated errors generated by merely mixing the numbers in the dataset. Our conclusions were based upon a particular case, which might led to wrong conclusions. To produce statistically reliable results, we decided to perform a Monte Carlo simulation to generate different realizations of artificial errors, and apply the method under consideration a number of times. All conclusions will arise only after consideration of a (large) number of cases. In 5.1.2 we will consider further how to obtain such realizations of artificial errors, and we will consider now how to evaluate the results.

Since the usefulness of the methods should be considered from either the user's or the data producers point of view, we have simulated within the codes the process of error detection - correction and further detection. We iterate as much as necessary in this way until some finishing criteria is satisfied. In the real-world example (described in *papers I and II)* we stopped when the Type I error was too high. In the simulation we stopped when a prescribed amount of points have been corrected or checked.

In order to to simulate a real case of correcting errors, we have provided a number of figures to analyze the output. It is assumed that once a value is pointed out as dubious it can be corrected, which in the statistical literature is known as the "perfect inspector" hypothesis. In the experiment, we seed the dataset with outliers, apply the methods, and continue until the finishing criteria is satisfied. Since in the real case the process is also iterative, we provide some intermediate measure of success. In bold we emphasize those measures of success that might be calculated during the depuration process. Among them:

a) RMSE of the remaining population (one possible measure of the accuracy)
b) **RMSE found up to the step**
c) Average of the absolute error of the remaining population (another measure of accuracy usually denoted as MAD in the literature; D stands for deviation).
d) **Average of the absolute error found up to the step**
e) **Type I** and Type II error up to the current step based on individual errors
f) **Type I** and Type II error up to the current step based on events

The others require knowing all the errors in advance. In order to allow a direct comparison between methods we will now introduce the concept of effort. We will define the effort as the ratio of the already checked or corrected values divided by the total number of values. So 100 per cent effort is the theoretical limit to find all errors (provided the inspector is perfect!).

Notice that all the measures mentioned are functions (not numbers!) related to the effort; this makes it difficult to qualify a method as better than another. Moreover, depending on the goals of the user, the criteria may be different. One user might want to locate as many errors as possible (disregarding their size), which corresponds to minimizing the Type II error for a given effort, while others might prefer to pick the largest errors in early stages, leading to minimizing either the RMSE or the MAD for a given effort.

So a more comprehensive statistics was sought, which should summarize the results. We devised one possible solution by considering that there exist two extreme possibilities for the methods: the *best method* will render only errors when requested and giving the larger errors first, and the *worst method* which will hide the errors as much as possible. When the process goes on since there is no possibility to repeat an error, the best method might be out of candidates, while the worst method will still hide them up to the end. The larger errors will be suggested at the end, as presented in fig. 5.9.

Any other method should operate outside the dashed areas: the top boundary indicated in fig. 5.9 is the worst possible outlier detection strategy, while the lower boundary is the best one. Notice that the best method yields accuracy 0 for any effort over N (equal to the number of errors in the dataset in per cent) impliying that there are no more errors in the set. The worst method in turn, will not modify the accuracy up to an effort 100-N, and it will diminish slowly from there on. At 100 per cent effort both lines end at 0, implying maximum accuracy. The line in the middle represents a possible real operation line; under the perfect inspector hypothesis it should be a non-increasing function, starting and ending in the same points than both the perfect and worst methods.

For a real operation example we will show only a fraction of the effort axis (see figs. 5.10 and 5.11). Notice that ascertain *which* are the larger errors depends if the method is event-oriented or datum-oriented. In the first case, the worst event is what contributed mostly to the error; however, it might not contain the worst *individual* errors.

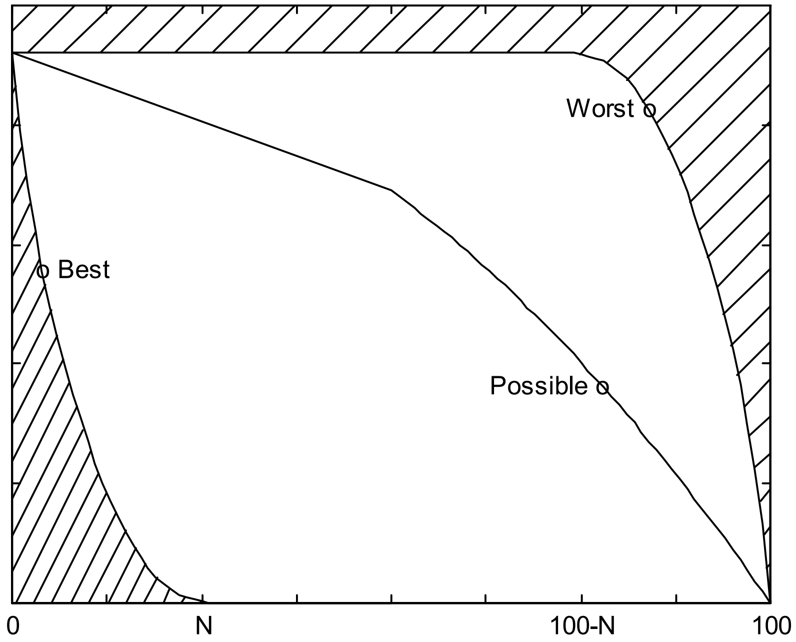

Figure 5.9 Sketch of the best, worst and a possible valid evolution of the accuracy in terms of the effort. N stands here for the fraction of the dataset which is wrong and is equal to the initial Type II error.

Since we are able to define a perfect and a worst method, and we know that any real method should lie in between, we might try to calculate some relative distance to the best operation curve considering also the worst one. We did so, and we define the following distance index function

$$I(effort) = \frac{\int_0^{effort} (worst(s) - curve(s)).ds}{\int_0^{effort} (worst(s) - best(s)).ds} \qquad (1)$$

which should have a value between 0 and 1, being preferred the larger values. Using the integral distance has some advantages over other alternatives, like to evaluate the statistics at a given effort. In figure 5.10 we show the best, the worst, and three possible operation lines.

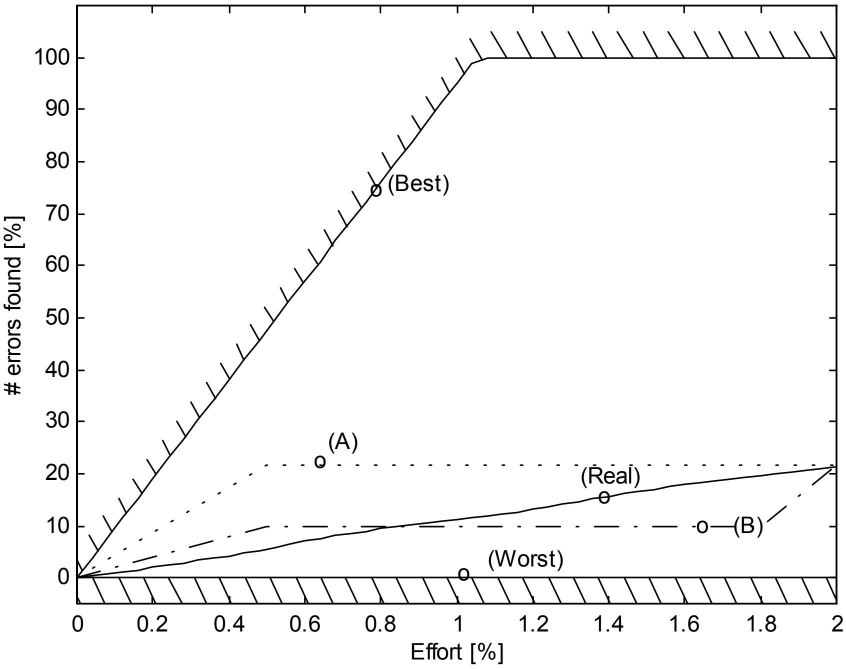

Figure 5.10 Example of best, worst, real and two ideal operation curves up to an effort of 2 per cent (data value-oriented) for the total numbers of errors found. The y-axis is related with the Type II error, while the slope is higher for lower values of the Type I error. Dashed areas indicate the limit of possible operation curves. The real curve (continuous), the (A) curve and the (B) curve are all valid examples of possible operation curves. See the text for an explanation.

The last three end at the same value for an effort equal 2 per cent, but the curve labelled (A) finds mostly errors in early stages, while curve labelled (B) did so at the end. Using simply the value of the function at 2 per cent effort will led to the (wrong) conclusion that all three alternatives are equally good. Summing up, since we will prefer curves which are closer to the best one we suggest to use the integral distance instead of the mere values at a given effort.

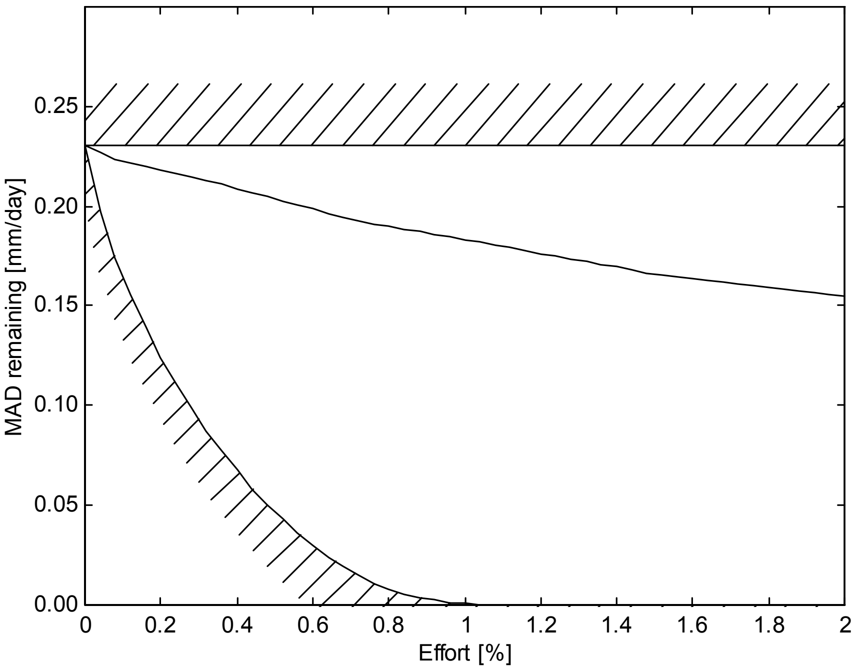

Figure 5.11 Example of best (at the bottom), worst (at top) and real (in the middle) operation curves up to an effort of 2 per cent (data value-oriented) in terms of the accuracy based on MAD (mean absolute deviation). Notice that for an effort of 1 per cent and over the best curve reports a MAD of 0.0 mm/day which is the maximum accuracy. Dashed areas indicate the limit of possible operation curves.

Figure 5.10 deserves some more comments, because it includes implicitly information on both Type I and II errors. Type I error is defined as the probability of misclassify a good value as wrong. Let N denote the initial fraction (in per cent) of the outliers in relation to the population. The function represented in fig. 5.10 is denoted as f(x), x being the effort (in per cent). The Type I error can be calculated for all x as

$$e_I = 1 - \frac{df}{dx}\frac{N}{100} \tag{2}$$

and for the Type II error the relationship is

$$e_{II} = \left(\frac{100 - f}{100}\right)\frac{N}{100} \tag{3}$$

so steeper functions f(x) will be preferred in order to decrease the Type I error. It can be shown that the slope of f(x) is strictly bounded with 100/N for any effort because at most we can find as many errors as candidates.

In order to allow a simple comparison among methods, we integrated all the operation curves to a prescribed effort level. The limit is different for *event-oriented* and *value-oriented* methods. We made the calculations up to an effort of 10 per cent in the first case, and up to 2 per cent in the second, which is roughly twice the number of values contaminated in the simulation.

In previous work we used the Type I error to decide whether to continue or not, based on the argument that it is one of the few statistics which can be calculated in the real case (either for the end user or the data producer). The other possibilities are the RMS and MAD of the absolute errors already found.

5.1.2 The error mechanism model

Before going into the results, it is suitable to analyze a delicate problem: how to simulate real errors. In order to make a Monte Carlo simulation, a tool to generate appropriate realizations of the typical errors is needed. The literature in this topic is scarce, and we found little guidance. One possible method is to simply mix the numbers in the table (Mixed Completely At Random, MIXCAR hereinafter). This procedure has the nice property of preserving the statistical characteristics of the population, but not of the time series. Provided all columns have similar pdf's the method might give acceptable errors. However, they might have little or no resemblance to real errors found in practice.

As part of the research we conducted an extensive depuration of the dataset using the outlier detection method described in *paper I* and denoted hereinafter as pcacov. After checking all candidates against paper records, we obtained a population of pairs *truth vs. wrong*, and we analyze some of its properties. It is clear that we found only those errors which are prone to be detected by the method, but we might left others in the dataset. If we plan to compare by Monte Carlo simulation the ability of different methods to detect errors, the decision of simulating errors like the ones the pcacov method certainly found will easily led us to the conclusion that such method is the best. The immediate question is how strongly depends the errors found population on the pcacov method.

In order to obtain a different error set independent of our procedure we typed twice a one year dataset. The file obtained were compared with the original one, and any discrepancy were analyzed; we concluded that pcacov detected almost all existing errors, so we attempted to simulate the errors located for the whole period. In *paper IX* we proposed an heuristics which produced errors clearly closer to the real ones than the MIXCAR criteria, and they were used in the final Monte Carlo experiment for the case of daily precipitation records. In *paper VII* we compared MIXCAR with errors detected by pcacov for the case of wind records, and we confirmed that the mere mixing is not good enough when compared with observed errors.

5.1.3 Results

The Monte Carlo experiment require running a number of simulations using the outlier generation mechanism quoted above, calculate the statistics and summarize the results. They are presented in different tables for event-oriented and datum-oriented methods. Table 5.1 shows the average values of the three different indices which are valid for event-oriented methods. The artificial neural network (ANN) based method named bp14 performs very satisfactorily in terms of finding errors (irrespective of their size). The others based on robust estimations of the Mahalanobis distance (MVE, MCD) as well as the one due to Hawkins (1974) show a similar performance. Once considering the size of the errors the picture changes a bit. On average, Hawkins74 is the best for MAD and RMSE, closely followed by ours, presented in *paper I* and denoted here as pcacov. The method based upon cross validation has a lower performance in all the indices, but close to the already mentioned methods.

Table 5.1 Average and probability (in per cent) of been the best of the Index-values after 450 runs for event-oriented methods. All indices are dimensionless

Method	Found vs. effort		Accuracy as MAD		Accuracy as RMSE	
	avg	best	avg	best	avg	best
bp1	52.3297	0.0	71.7055	0.0	63.7345	0.2
bp7	58.4470	3.6	75.2580	0.7	64.1095	0.7
bp14	**59.3301**	44.2	**76.0892**	4.0	64.9154	0.7
crossva05	54.2718	0.0	73.4323	10.4	64.6805	16.7
pcacov	56.0469	0.2	75.5096	6.9	**65.5949**	10.4
Hawkins74	58.4829	20.4	**77.7446**	70.2	**68.6106**	68.7
MVE	**59.0541**	12.7	**76.3609**	7.8	**65.6407**	2.7
MCD	**59.1978**	18.9	75.3159	0.0	63.8466	0.0
Rocke96	49.5981	0.0	55.2424	0.0	34.7747	0.0
Hadi94	55.3678	0.0	44.3686	0.0	30.9342	0.0

The mean value of the index might not give a correct picture, so we attempted to compare all methods to all methods for each index. This has been done in table 5.2 for the errors found vs. effort curve, in table 5.3 for the MAD index and in table 5.4 for the RMSE index. The entry (i,j) of the table is the probability estimate that method i will produce an index lager than method j. The three best options are presented in boldface.

Table 5.2 Probability estimate (in per cent) that the Index for errors found vs. effort for method i exceeds those of method j for event-oriented methods. The last row is the probability of not being the best option. Results after 450 runs.

	bp1	bp7	bp14	crossva05	pcacov	Hawkins74	MVE	MCD	Rocke96	Hadi94
bp1		0.0	**0.0**	14.0	0.9	0.0	**0.0**	**0.0**	96.2	2.4
bp7	100.0		**9.8**	97.3	97.6	47.1	**18.9**	**12.2**	100.0	98.2
bp14	100.0	90.2		100.0	99.6	73.8	**72.7**	**62.9**	100.0	99.8
crossva05	86.0	2.7	**0.0**		67.6	2.0	**0.4**	**0.0**	86.0	74.0
pcacov	99.1	2.4	**0.4**	32.4		0.7	**0.7**	**0.7**	100.0	65.8
Hawkins74	100.0	52.9	**26.2**	98.0	99.3		**31.1**	**33.6**	100.0	96.2
MVE	100.0	81.1	**27.3**	99.6	99.3	68.9		**34.9**	100.0	100.0
MCD	100.0	87.8	**37.1**	100.0	99.3	66.4	**65.1**		100.0	100.0
Rocke96	3.8	0.0	**0.0**	14.0	0.0	0.0	**0.0**	**0.0**		0.0
Hadi94	97.6	1.8	**0.2**	26.0	34.2	3.8	**0.0**	**0.0**	100.0	
Average	87.39	35.43	**11.22**	64.59	66.42	29.19	**20.99**	**16.03**	98.02	70.71

Table 5.3 Probability estimate (in per cent) that the Index for MAD vs. effort for method i exceeds those of method j for event-oriented methods. The last row is the probability of not being the best option. Results after 450 runs.

	bp1	bp7	bp14	crossva05	pcacov	Hawkins74	MVE	MCD	Rocke96	Hadi94
bp1		0.7	**0.0**	14.0	2.4	**0.0**	**0.0**	0.0	100.0	100.0
bp7	99.3		**20.9**	27.6	42.9	**12.7**	**15.1**	48.2	100.0	100.0
bp14	100.0	79.1		45.3	60.2	**23.1**	**37.3**	89.6	100.0	100.0
crossva05	86.0	72.4	**54.7**		59.6	**21.8**	**47.8**	78.0	92.0	100.0
pcacov	97.6	57.1	**39.8**	40.4		**17.6**	**34.0**	59.1	100.0	100.0
Hawkins74	100.0	87.3	**76.9**	78.2	82.4		**74.7**	84.0	100.0	100.0
MVE	100.0	84.9	**62.7**	52.2	66.0	**25.3**		99.1	100.0	100.0
MCD	100.0	51.8	**10.4**	22.0	40.9	**16.0**	**0.9**		100.0	100.0
Rocke96	0.0	0.0	**0.0**	8.0	0.0	**0.0**	**0.0**	0.0		100.0
Hadi94	0.0	0.0	**0.0**	0.0	0.0	**0.0**	**0.0**	0.0	0.0	
Average	75.88	48.14	**29.49**	31.97	39.38	**12.94**	**23.31**	50.89	88	100.00

The bp14 method (based upon ANN) is the best in order to detect errors. However, it is not the best for detect the larger errors. This make evident the high sensitivity of the ANN to errors, possibly due to the nonlinear characteristics of the transfer function. In opposition, the crossvalidation method performs better in terms of RMSE and MAD rather than the first index. Our proposed method based on PCA have clearly a role to play. It is interesting to notice that among the Mahalanobis-distance based methods, the method named Hawkins74 typically outperforms all standard procedures in terms of all the indices. In addition, it requires only a fraction of time to calculate its parameters (which depends only on the sample's covariance matrix). If we consider CPU requirements, Hawkins74 is the best option; it balance a fairly good performance in all indices with a comparatively low demand on system. All considered ANN were heavy to train but rather cheap to use, while the crossvalidation method is more expensive to use. Numbers about floating point operations and time required for use each method were also collected during the experiments, and they are summarized in López *et al.* 1997.

Table 5.4 Probability estimate (in per cent) that the Index for RMSE vs. effort for method i exceeds those of method j for event-oriented methods. The last row is the probability of not being the best option. Results after 450 runs

	bp1	bp7	bp14	crossva05	pcacov	Hawkins74	MVE	MCD	Rocke96	Hadi94
bp1		40.4	23.8	**16.2**	23.8	**14.0**	**9.1**	46.7	100.0	100.0
bp7	59.6		41.6	**18.0**	34.7	**15.3**	**31.3**	57.3	100.0	100.0
bp14	76.2	58.4		**22.0**	38.2	**22.7**	**27.6**	87.8	100.0	100.0
crossva05	83.8	82.0	78.0		59.6	**26.7**	**71.6**	84.9	100.0	100.0
pcacov	76.2	65.3	61.8	**40.4**		**24.7**	**52.0**	76.4	100.0	100.0
Hawkins74	86.0	84.7	77.3	**73.3**	75.3		**74.2**	81.6	100.0	100.0
MVE	90.9	68.7	72.4	**28.4**	48.0	**25.8**		99.6	100.0	100.0
MCD	53.3	42.7	12.2	**15.1**	23.6	**18.4**	**0.4**		100.0	100.0
Rocke96	0.0	0.0	0.0	**0.0**	0.0	**0.0**	**0.0**	0.0		96.4
Hadi94	0.0	0.0	0.0	**0.0**	0.0	**0.0**	**0.0**	0.0	3.6	
Average	58.44	49.13	40.79	**23.71**	33.69	**16.40**	**29.58**	59.37	89.29	99.60

The method Rocke96 is based upon a high breakdown estimate of the covariance matrix as described by Rocke and Woodruff (1996). We found no satisfactory explanation for its poor performance. One possible explanation is related to the present limitations of the program. The code (available in statlib) presently cannot handle events with repeated values. Rocke (1997) suggested to use a slightly randomly perturbed database, and we added to all readings uniformly distributed perturbations at most of absolute size 0.01 mm/day (one tenth of the data resolution). This might significantly affect the estimators. Another factor to explain the relatively poor behavior is the significant deviation of daily precipitation records from a gaussian pdf (an argument also raised by Hadi (1997), a situation not considered in the theory behind these methods.

As mentioned before, most of the work in multivariate outlier detection in statistics is event rather than datum-oriented. Also, most experiments and results are based on small datasets (less than 30 events) and a limited number of variables. When faced with the less manageable thousands of events typically handled in meteorology we cannot merely detect the event, and there is a need to pinpoint the outlying value(s) within the event. There are also other practical reasons: in our case the precipitation information is compiled in books by station, so it is cumbersome to check routinely all the values for a given event, because it requires to handle 10 books at the same time. Fortunately, some of the methods outlined can be tailored to produce a narrower list of candidates within each event event (using a procedure first suggested in *paper I* and described in 4.1.3). The list includes all Mahalanobis-based methods (Hadi94, Rocke96, Hawkins74, MCD and MVE) as well as our proposed pcacov.

In table 5.5 we compare the results of the simulations of the Mahalanobis-based methods with and without our procedure for detecting outliers within the event. The alternatives were to use only the unlikely readings of the event, or all of them. It is shown that for Hawkins74, MCD, MVE as well as our proposed pcacov methods the sensitivity based approach improves the results, while for Hadi94 and Rocke96 it is only of marginal importance.

Table 5.5 Probability estimate (in per cent) that the three different index improves after adding a sensitivity analysis to the standard event-oriented method for all Mahalanobis-like oriented methods. Results after 450 runs.

	MVE	MCD	Hadi94	Rocke96	Hawkins74
Index	datum vs. event	datum vs. event	datum vs. event	datum vs. event	datum vs. event
Found vs. effort	100.0	100.0	70.7	62.9	95.6
MAD	100.0	100.0	44.4	32.4	100.0
RMSE	100.0	98.0	37.8	9.3	100.0

In table 5.6 the first results for the datum-oriented methods shows (surprisingly!) that the MVE plus our sensitivity approach outperforms the others in terms of the number of errors found and is the second in terms of both accuracy as MAD and RMSE. The MCD method (with similar CPU requirements as MVE) shows a similar behavior, and it is good for detecting errors. For the accuracy as MAD and RMSE our well known Hawkins74 is the most effective. Again our proposed pcacov method show better performance for MAD and RMSE (i.e. larger size errors) and not so good for the first index.

Table 5.6 Average and probability (in per cent) of been the best of the Index-values after 450 runs for datum-oriented methods. All indices are dimensionless

Method	Found vs. effort		Accuracy as MAD		Accuracy as RMSE	
	avg	best	avg	best	avg	best
MVE	**17.0613**	96.0	**27.5818**	0.9	**21.7291**	3.8
MCD	**15.3924**	0.9	22.6277	0.0	16.5569	0.0
Hadi94	12.1733	0.0	14.8600	0.0	9.0992	0.0
Rocke96	11.2805	0.0	12.6134	0.0	6.7587	0.0
pcacov	12.1098	0.0	**26.6347**	2.9	**19.2210**	2.9
Hawkins74	**15.0240**	3.1	**33.0529**	96.2	**25.9581**	93.3

The equivalent numbers from table 5.1 cannot be compared with those of table 5.6 for two reasons. The first one is that the best and worst operation curves are different for event and datum oriented process, despite that they can be represented in the same axis. They will coincide if and only if there is at most one error per event. If there is more than one, for the event-oriented case the best option is to choose as the first candidate the event which contributed most to the error measure, while for the datum oriented case the option will be to choose the largest error. The second reason is that, despite their similar name, the datum-oriented methods are in fact different than those reported. In addition, for table 5.1 all indices have been integrated up to an effort of 10 per cent, while for table 5.6 we integrated only up to 2 per cent.

In tables 5.7 to 5.9 the results show that the Mahalanobis-distance based method using the MVE as a kernel plus the our proposed sensitivity leds to the best results for merely find the errors, but when error size is taken into account, Hawkins74 performs better for both the accuracy measures considered.

Table 5.7 Probability estimate (in per cent) that the Index for errors found vs. effort for method i exceeds those of method j for datum-oriented methods. The last row is the probability of not being the best option. Results after 450 runs

	MVE	MCD	Hadi94	Rocke96	pcacov	Hawkins74
MVE		**99.1**	100.0	100.0	100.0	**96.9**
MCD	**0.9**		99.8	100.0	99.8	**61.1**
Hadi94	**0.0**	**0.2**		86.9	48.9	**2.4**
Rocke96	**0.0**	**0.0**	13.1		26.0	**0.7**
pcacov	**0.0**	**0.2**	51.1	74.0		**2.0**
Hawkins74	**3.1**	**38.9**	97.6	99.3	98.0	
Average	**0.80**	**27.68**	72.32	92.04	74.54	**32.62**

Table 5.8 Probability estimate (in per cent) that the Index for MAD vs. effort for method i exceeds those of method j for datum-oriented methods. The last row is the probability of not being the best option. Results after 450 runs

	MVE	MCD	Hadi94	Rocke96	pcacov	Hawkins74
MVE		100.0	100.0	100.0	**59.3**	**1.1**
MCD	**0.0**		100.0	100.0	**9.3**	**0.0**
Hadi94	**0.0**	0.0		96.7	**0.0**	**0.0**
Rocke96	**0.0**	0.0	3.3		**0.0**	**0.0**
pcacov	**40.7**	90.7	100.0	100.0		**2.9**
Hawkins74	**98.9**	100.0	100.0	100.0	**97.1**	
Average	**27.92**	58.14	80.66	99.34	**33.14**	**0.80**

Table 5.9 Probability estimate (in per cent) that the Index for RMSE vs. effort for method i exceeds those of method j for datum-oriented methods. The last row is the probability of not being the best option. Results after 450 runs

	MVE	MCD	Hadi94	Rocke96	pcacov	Hawkins74
MVE		100.0	100.0	100.0	**78.2**	**4.0**
MCD	**0.0**		100.0	100.0	**18.9**	**0.0**
Hadi94	**0.0**	0.0		94.2	**0.4**	**0.0**
Rocke96	**0.0**	0.0	5.8		**0.0**	**0.0**
pcacov	**21.8**	81.1	99.6	100.0		**3.1**
Hawkins74	**96.0**	100.0	100.0	100.0	**96.9**	
Average	**23.56**	56.22	81.08	98.84	**38.88**	**1.42**

Just to give an idea about the variability observed in the indices and other measures of success, we provide some plots with the observed histograms. In figure 5.12 three histograms corresponding to three different effort levels are sketched for the case of the Type I error using the ANN bp14 working as event-oriented detection tool. It is clear that the type I error increases as soon as the process goes on, because most of the obvious errors are found in the early stages of the procedure.

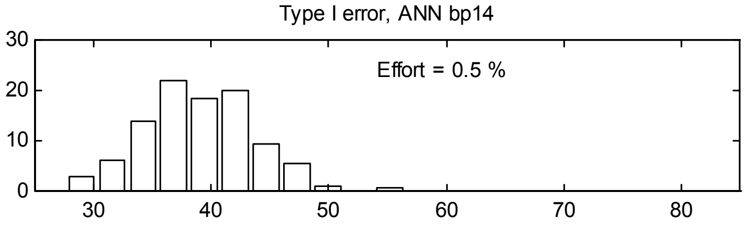

Figure 5.12 Evolution of the Type I error histogram for different effort levels, for the ANN bp14 (event-oriented method). Results after 450 simulations

In figure 5.13 we presented the observed histogram for the three Indexes (found vs. effort, accuracy as MAD and as RMSE) for the method Hawkins74 + sensitivity (datum-oriented). The plot can be considered typical for other methods also. In figure 5.14 the histograms for the RMSE found up to a given effort for the same method confirm our previous comment that larger errors are more prone to be found in early stages of the depuration process. All of the results have been obtained after a Monte Carlo simulation of 450 replications. In order to check that a stationary situation has been reached we applied the Kolmogorov-Smirnov test (Koroliuk, 1986) using a confidence level of 5 per cent.

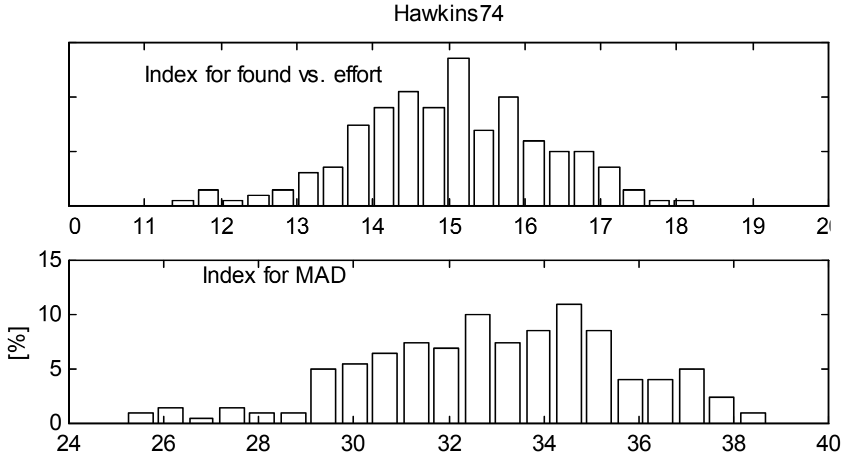

Figure 5.13 Histogram of the three Index considered after 200 simulations using the Hawkins74 plus sensitivity method (datum-oriented)

5.1.4 Discussion

There are some interesting points to consider. One is the good performance of the Hawkins (1974) method both for event-oriented and datum-oriented case. The difference compared with the MVE is important for the Index for the found vs. effort curve, but of less importance for those based upon measures of the accuracy in the case of datum-oriented. Calculating the exact MVE is a heavy task; state-of-the-art code will require ages to find it (Hawkins 1997) so we used some well established, public domain software and limited the random trials. The requirements for the Hawkins74 code are modest: it requires calculation of the sample covariance matrix, its eigenvectors and the eigenvalues. After a simple analysis of the loadings in the eigenvectors the number of scores to retain can be determined (which in turn is considered the limit between the noisy and structural information of the cloud) and an open region in the R^w space can be defined for a given confidence level. Any event represented in R^w space with a point inside the region is considered outlier-free.

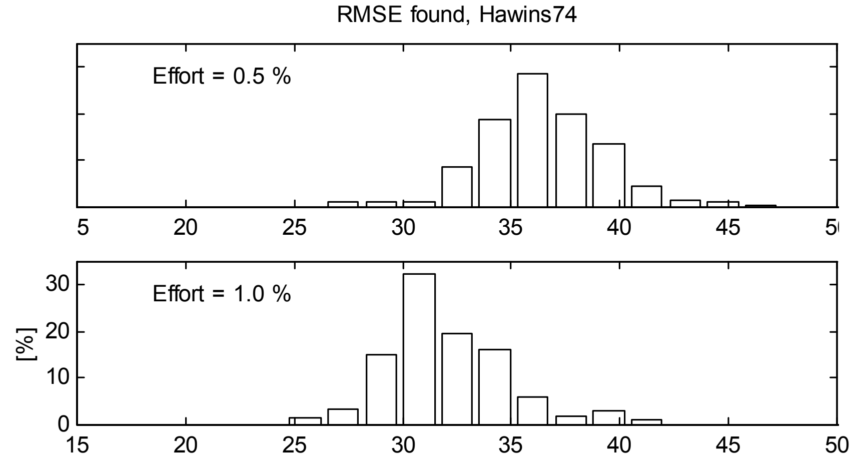

Figure 5.14 Histogram of the RMSE found up to a given effort, for three different efforts, using the Hawkins74 method+sensitivity. Results after 200 simulations.

It is interesting to note that, as implemented, the method Hawkins74 has a low breakdown. The definition of breakdown considers completely arbitrary errors, and due to the nearly-real error generation mechanism we do not allow them to exist. Low breakdown imply that large errors might severely affect both the estimate of the location of the cloud (i.e. its centroid) as well as the covariance matrix, which has not been noticed for this method. Since the dataset is completely dominated by zeros, even with gross errors the estimate of the location of the cloud might be only slightly displaced from its "robust" location (close to the origin in R^w). However, the same argument do not hold for the estimate of the covariance matrix. An open question is why, even using a non robust estimator of the covariance, the rule for separate the noise from the structure of the cloud in terms of outliers seems robust.

The use of regions is also in line with procedures like MVE and MCD; but such regions are bounded in R^w. The method proposed in *paper I* also considers an open region in R^w space, but slightly different from the one of Hawkins74; it requires that the noisy scores belong to prescribed intervals, which can be represented as some hypercylinder with hyperrectangular cross sections, while Hawkins74's cross section are hyperellipses.

We proposed the application of ANN to this problem. We are not aware of any published successful interpretation of the intermediate stimuli within the ANN, because it is claimed that the non-linearity precludes further interpretation. We gave arguments to consider our PCA-based method (pcacov) as a particular ANN, with one hidden layer of linear neurons, one with rectangular window transfer functions and the output working as a logical OR operator. For the outlier detection problem all the ANN considered have been trained with supervised learning, with the imputation of missing values as the objective. In the general case, the output of the first hidden layer has been considered as a point in R^H space, being H the number of neurons. Some rule has been proposed in order to distinguish between the structural and the noisy neurons, the latter being those which are activated only in unusual cases. For those neurons some intervals for their outputs have been proposed, and any event which produces a stimulus out of range for any of the noisy neurons is pointed out as outlying. The best results have been obtained with a rather simple ANN, using our proposed transfer functions instead of the traditional tansig. Considering the crude reasoning used the results are more than satisfactory. ANN are, as MVE and MCD, heavy in CPU requirements. Brute force has been applied in order to train it, so it is possible that these preliminary figures can be improved.

Two other high breakdown methods were considered. The one proposed by Hadi (1994) gave satisfactory results for event-oriented outlier detection in terms of Type I and II errors (both considered in the errors found vs. effort index). Such performance appears to be independent to the size of the errors, as arises from its poor performance when considering accuracy measures (i.e. MAD vs. effort and RMSE vs. effort indices). The code from Rocke and Woodruff (1996) outperformed Hadi's method in terms of accuracy, but is worse in terms of the first index. There might be a number of reasons for this: a) the code presently did not handle events with exactly the same readings, and we needed to perturb the dataset slightly in order to proceed. Another reason (mentioned independently by Hadi and Rocke after personal communications) is that both methods assume multivariate gaussian pdf. One possible explanation is that these methods might be more sensitive to deviations from a gaussian pdf than the other procedures considered.

All procedures degrade significantly when dealing with the datum-oriented problem. For those methods based upon Mahalanobis-like distance we devised a sensitivity approach which produced better results. However, the compound method is no longer affine equivariant, which implies that after performing a linear transformation of the dataset the candidates might be different.

The problem of generating random realizations of the dataset with feasible outliers is not a minor one. We analyzed the outputs of the original paper correction process, being the errors detected using the pcacov method (all described in *paper I*). We carried out a new, independent re-typing for one year of data in order to have a different error detection and correction mechanism. It cannot be assured that the resulting dataset for that year is error free, but it is felt that it is very close. The statistical properties of the errors observed in such year were similar to those of the overall period. After a qualitative analysis of the errors, some simple error-generation mechanism were postulated, and we attempted to model errors using it. The agreement with errors observed in practice was good, as reported in *paper IX*. The pcacov procedure was successfully applied also to surface wind *(paper VII)*, and surface presure with comparable results.

5.1.5 Conclusions

We have conducted an extensive experiment, the results of which supersede our previous results presented in *paper I*. We have included the state-of-the-art algorithms for the event-oriented multivariate outlier detection problem, others not longer considered after publication more than 20 years ago, and suggested new ones. We presented the results by using concepts of distances of functions, which we believe is a major improvement over our previous results, and appropriate to handle large datasets as the one considered.

For practical reasons, most of the literature in statistics concentrates on small datasets, of a few dozens of events and with only few variables. One possible reason is editorial. Other is the heavy CPU requirements for medium to large size datasets which precludes extensive analysis. The work of Rocke and Woodruff (1996) is an exception of that rule. We are not aware of any paper published in the meteorological literature about an experiment like the one reported here. Large meteorological datasets are checked with procedures relying either on expert rules and/or using auxiliary mathematical models, being both variable-dependent. In additon, they require substantial CPU resources and independent data.

5.2 Missing value problem in tabular data: the daily precipitation example

5.2.1 Experiment strategy

In order to make an objective comparison among different methods we decided to simulate randomly located missing values, and apply the different proposed methods to fill the gaps. In order to assess whether or not the missing value pattern depends on the observed variables we applied the test for MCAR (Missing Completely At Random) described by Little (1988). We ran the test both with the original and the simulated holes, and in both cases the missing value pattern is independent from the observed as well as the unobserved values.

After applying the imputation methods some statistics are calculated attempting to measure the discrepancies among the imputated and the original value. In *paper II* we tested four methods, namely: nearest neighbor, ordinary time interpolation and two new based on PCA analysis and they were applied to one dataset with randomly generated missing values. The best results were obtained with one of the new methods.

Further research made evident the need for generate multiple realizations of the datasets with missing values in order to derive statistically valid conclusions. The overall process was processed by a Monte Carlo procedure. Also we include in the test many other methods, both linear and non linear, in order to make a more comprehensive comparison. All has been accomplished within a research project (to be described in full in López *et al.* 1997) and partially reported here; *paper III* can be regarded as a progress report on the topic.

5.2.2 Results

The results obtained after 250 simulations[1] are summarized in tables 5.10 and 5.11. It should be noticed that only slightly improved results have been obtained by those methods which use information from two days (*bp12, bp17*, *gandin4, gandin6 and gandin_diario, etc.*). Among those which use only information of a single day, the best results are obtained by the minimum 95 percentile, closely followed by the Ordinary Least Squares method.

In 4.1.5 we introduced the ANN for quality control, but the different architectures used have not been presented. They are summarized in Table 5.12 where the prefix "bp" stands for backpropagation network. Under the heading "layers" the type and

[1] The figures differ to some extent from those presented in the paper, because they were derived using an intermediate version of the dataset (prior finishing the real error´s correction task)

number of neurons is presented. *Sinh* and *Asinh* stands for the hyperbolic sin and its inverse, and the other types are denoted according to Demuth and Beale (1994).

Table 5.10 Preliminary results in mm/day for the different imputation methods which use and predict data for the same day. The expected value and the 75, 85 and 95 percentile of the distribution of the absolute error, and its RMS are presented and compared. A qualitative indication of the required resources is provided. In connection with Table 5.11 the five best results for each estimator have been higlighted in bold.

Algorithm	Average	75 per cent	85 per cent	95 per cent	RMSE	Resources	
	mm/day	mm/day	mm/day	mm/day	mm/day	High	Low
bp1	2.45	1.88	4.20	12.17	6.54	*	
bp2	2.63	2.23	4.35	12.50	6.67	*	
bp7	2.74	1.83	3.57	12.23	7.47	*	
bp10	2.44	1.40	4.14	13.52	7.16	*	
bp14	2.37	1.55	3.99	12.27	6.53	*	
bp22	3.29	2.01	5.80	18.38	9.92	*	
bp23	2.81	0.44	**3.09**	18.08	9.83	*	
cressman	2.21	1.18	3.88	12.53	6.62		*
daymean	2.42	1.61	4.42	13.28	6.92		*
expert's distance	2.44	**0.00**	3.85	15.25	8.02		*
gandin	2.25	1.21	3.76	12.03	6.35		*
gandin20	2.28	1.29	3.78	**11.96**	**6.34**		*
gandin3a	2.42	1.38	4.12	13.51	7.16		*
gandin5	2.11	1.11	3.72	12.05	6.35		*
gandin7	**1.97**	0.42	**2.89**	**11.90**	6.52		*
gandintrans	2.82	2.37	4.60	13.13	7.51		*
geometrical distance	2.41	**0.00**	3.81	15.03	7.94		*
hotdeck	2.77	**0.40**	4.37	16.84	8.43	*	
kulback	2.83	**0.06**	4.73	17.62	9.06		*
lms	2.19	1.27	3.90	12.22	6.51	*	
lts	**2.06**	0.89	3.31	**11.88**	6.50	*	
POPS	2.37	1.35	3.80	12.03	6.42		*
least squares	**2.10**	0.52	3.63	11.96	**6.33**		*
least 95 percentile	**2.09**	0.60	3.68	**11.81**	**6.29**	*	
least average	**2.03**	0.50	**3.29**	11.94	6.44	*	
modal value	2.95	**0.00**	**2.79**	20.24	10.41		*
station average	4.95	3.06	**3.24**	17.28	9.99		*

The number of inputs are the 9 readings from the current day, except for bp12 and bp17 (which in addition uses 10 readings from the day before) and bp11, which only uses the readings from the day before. The results for the ANN were somewhat poor. Among others, the reasons might be: a) inapropriate architecture and/or training procedures b) the training algorithm is directed towards minimizing the sum of squares; a maximum likelihood approach would have been more suitable and c) the precipitation field might not be smooth enough.

It should be stressed that, since the database used in *paper III* had many errors, it is possible that the methods suggest suitable values and the outliers affect some of the considered statistics. This is unlikely to occur for the 85, 95, etc. percentile, and that might be an explanation of the different ranking observed for the ANN denoted bp7 in the paper and in the calculations here presented. The present results were based on a slightly different dataset because we continued correcting errors after those preliminary calculations.

Table 5.11 Preliminary results in mm/day for the different imputation methods. The table summarizes methods which use data from the same day, the day before or both. The expected value and the 75, 85 and 95 percentile of the distribution of the absolute error, and its RMS are presented and compared. A qualitative indication of the required resources is also provided. In connection with Table 5.10 the five best results for each estimator have been higlighted in bold.

Algorithm	Average	75 per cent	85 per cent	95 per cent	RMSE	Resources	
	mm/day	mm/day	mm/day	mm/day	mm/day	High	Low
bp11	4.61	4.15	6.75	17.13	9.60	*	
bp12	2.75	2.45	4.51	12.25	6.82	*	
bp17	2.59	2.12	4.25	12.27	6.62	*	
gandin4	2.22	1.55	3.97	**11.97**	**6.30**		*
gandin6	2.35	1.66	4.06	12.03	**6.31**		*

The method denoted as *POPS* corresponds to the one proposed in *paper II*. It is one of the best options among the ones which use only data from the same day. It also requires modest CPU resources. The results from this simulation support only partially those discussed in *paper II*: for the case of daily precipitation records *POPS* outperformed the methods based on geometrical distance, but not those of Optimum Interpolation (gandin and gandin20).

5.2.3 Discussion

The results confirm that, despite the different methods used, the daily precipitation missing value problem is a very difficult one. The relative differences between the best values obtained and a reference method like nearest neighbor are low: the RMSE drops 20.6 per cent, the 95 percentile 20.9 per cent and the MAD 18.3 per cent.

All commonly used linear methods have been included in the test. For each measure of success (i.e. RMSE, p95 or MAD) there are specific choices of the coefficients oriented to make a linear combination the best. All other linear method should perform worse. In any case, even the "best-among-linear" methods have coefficients determined using only a fraction of the dataset with no missing values. However they were later applied to the whole population. Slight differences between both sets might explain why the Ordinary Least Squares is not optimal in terms of RMSE. For the Least 95 percentile and Least Average method we have solved nonlinear optimization problems and the algorithms might have been trapped in local minima, or the maximum number of iterations has been defined too low.

Table 5.12 Brief information about the architecture of the different artificial neural networks used in the missing value problem for the precipitation. f(precipitation) denotes the transformation which renders a nearly uniform probability density function. t and t-dt denote values from the day and the day before. ANN bp22 and bp23 have a variable number of neurons in the hidden layer. The input in all cases can be regarded as being scaled by its temporal average in order to make the weights and bias dimensionless

Our coded name	*Layers*	*Input variable* $^{(**)}$	*Using data from days*	
			t	t-dt
bp1	Tansig(6)/Purelin(1)	precipitation	X	
bp2	Tansig(6)/Purelin(1)	precipitation - daily mean	X	
bp7	Purelin(8)/Logsig(4)/Logsig(1)	f(precipitation)	X	
bp10	Tansig(6)/Purelin(1)	precipitation - first guess	X	
bp11	Tansig(6)/Purelin(1)	precipitation		X
bp22	Tansig(*)/Purelin(1)	precipitation	X	
bp23	Tansig(*)/Purelin(1)	sqrt(precipitation)	X	
bp14	Sinh(4)/Asinh(1)	precipitation	X	
bp12	Tansig(6)/Purelin(1)	precipitation	X	X
bp17	Sinh(4)/Asinh(1)	precipitation	X	X

In theory, Ordinary Least Squares and gandin20 should render exactly the same results if all coefficients were determined using the same data. This is not the case, because we used all the data available for the gandin20 procedure, and limited ourselves to events with no missing values for the OLS.

The high breakdown methods LTS and LMS produced results similar to those of OLS (a low breakdown one). We conclude that reason is that the dataset is almost free from gross errors. It is interesting to notice that our method POPS produced better results in terms of RMSE and is comparable in this respect to least 95 percentile, despite its very limited CPU requirements.

Once all the linear methods have been compared we should consider the non-linear ones. The author is not aware of applications of non linear methods in the field of meteorology, but they certainly exist in other fields. The ANN approach looks promising, despite its high CPU requirements. The somewhat different results obtained in the paper deserve some discussion. One explanation is that we trained the ANN using the original data (prior to manual correction) and tables 5.10 and 5.11 were derived using a corrected dataset. The figures were in all cases lower than those presented in *paper III*, but the ANN were more (unfortunately!) robust. It should be stressed that once the ANN is trained, its operational cost is very low. The training itself is a complex task. Despite we have used state-of-the-art software for the training, it easily get stuck in local minima failing to go to a global one. This has been proved while analyzing the effect of increasing the number of neurons in the hidden layer. The ANN should fit the data better but, in some cases, the training algorithm failed to do so. That poses a question mark about how much the already obtained results could be improved by using better training procedures. Some other posibilities are: a) design different ANN architectures, like non-fully connected nets; b) change the training routine in order to minimize other objective functions than RMSE. This might provide high breakdown abilities to the training similar to those described for LTS, LMS, etc.; c) use different neurons which might change weights with time.

5.2.5 Conclusions

Commonly applied methods based upon mere substitution by a neighbor or by a constant gave poor results. The least 95 percentile method wast (unexpectedly) better than optimum interpolation (gandin20) and ordinary least squares method in terms of RMSE, probably due to the different policy adopted for the original missing values existing. The ANN methods have been trained in order to minimize the RMSE, but the training was performed with early (contaminated) versions of the dataset. The training algorithms have shown problems to escape from local minima, leading to suboptimal results. Better training strategies are needed.

5.3 Missing value problem in tabular data: the hourly surface wind example

The methods used for this experiment were well described in 4.2. Some of them should not be used due to some peculiarities of the wind data. For instance the 0 value is no longer the most probable value, and events with all zero wind are unlikely. Another point is that wind data is not a scalar, but a vector. All the methods tested by us work with homogeneous data, thus direction and modulus cannot be straightforward combined in the same table. Therefore, we transform the wind data to its components along the E-W and N-S directions, with appropriate sign. This might make a difference with the precipitation data case.

Another point is that some methods which use the temporal correlation of the phenomena might be useful in this case, as opposed to the case of precipitation data. Among others, the TIPS (Time Interpolation of Principal Scores) method may render good results. The method has been proposed for the first time in *paper II*.

5.3.1 Results

In *paper VIII* we reported the results of the application of the methods for systematic and random location of missing values. Most of the dataset consists of hourly values. However, some data were only acquired three times a day, which motivated the use of methods for both systematic and random location of missing values. In table 5.13 the performance of five methods is shown; the fifth method is simply time interpolation of the individual station records, which results in a particular case of the TIPS method (the notation 1:10 has been used to consider as noisy all scores from Principal Component 1 to 10). Our proposed method TIPS gave the best results, while the standard Gandin's one falls somewhat below. Except for the case of few interpolated terms, the other results suggest that most of the information is contained in the first three PC as it was suggested in the paper. There we attempted to validate these results, removing values at random and applying the same methods again; the results seem to support the findings in table 5.13. The TIPS scheme outperforms the others, but the number of interpolated terms is somewhat higher. The good performance of the simple standard interpolation scheme suggest that the hourly wind field is oversampled in time, at least in Uruguay.

To test this, we extended our dataset to 6 years (but subsampled to 8 readings per day), and performed a Monte Carlo simulation in order to obtain more reliable results. In table 5.14a and 5.14b the preliminary results are shown. We have also included ANN methods, as well as the best linear estimators for sum of squares, MAD and 95 percentile. As before, we considered those methods that rely upon information from more than one event in a separate table. The mere time series interpolation has been included as a separate method, and two versions of our proposed TIPS are considered, depending on whether or not the individual variables are normalized to zero mean and unitary standard deviation. The number of uncontrolled scores was defined according to Hawkins (1974).

Table 5.13 RMSE and the mean of the error (not MAD!) of the difference between measured value and calculated value upon imputation assuming no missing values in Melo and readings only at 8, 14 and 20 hours local time for the other 4 NWS stations. Data from the year 1990-91. In boldface the most significant outputs. (taken from *paper VIII*)

TIPS			POPS		
Interpolated terms	RMSE (m/s)	Mean (m/s)	Penalized terms	RMSE (m/s)	Mean (m/s)
1:10	2.06	0.09625	10:10	3.41	0.10094
1:9	2.06	0.09669	9:10	3.41	0.10274
1:8	2.05	0.09613	8:10	3.39	0.10452
1:7	2.06	0.09671	7:10	3.28	0.06608
1:6	**2.05**	**0.08151**	6:10	3.26	0.06191
1:5	2.06	0.09585	5:10	3.23	0.04485
1:4	2.05	0.09541	4:10	3.21	0.01852
1:3	2.05	0.08414	3:10	3.40	0.01686
1:2	2.11	0.08763	2:10	2.97	0.00177
1:1	2.73	0.07331	**1:10**	**2.84**	**0.05171**

Results obtained assigning the mean value	3.24	0.28839
Results obtained with Gandin´s method	2.84	0.05353

Table 5.14a Preliminary results in m/s for the different imputation methods who uses and predicts data for the same day. The expected value and the median, 75, 85 and 95 percentile of the distribution of the absolute error, and its RMS are presented and compared. In bold the six best results for each estimator.

Algorithm	MAD m/s	Median m/s	75 per cent m/s	85 per cent m/s	95 per cent m/s	RMSE m/s
bp1	**2.26**	**1.42**	**2.94**	**4.17**	**7.37**	**3.50**
bp14	**2.36**	1.45	**3.03**	**4.38**	**8.02**	**3.68**
gandin	2.53	1.64	3.28	4.66	8.28	3.76
gandin20	2.52	1.64	3.27	4.65	**8.27**	3.75
gandin5	2.52	1.57	3.26	4.72	8.52	3.85
POPS	3.23	1.84	3.93	5.85	11.17	6.61
ordinary least squares	2.52	1.56	3.27	4.74	8.53	3.84
least 95 percentile	2.52	1.57	3.26	4.72	8.50	3.84
least average	2.46	1.45	3.14	4.64	8.76	3.89
staverage	3.24	2.36	4.43	5.74	10.21	4.53

5.3.2 Discussion

We have applied a number of alternative methods to a surface wind dataset. Some of these methods have also been sucessfully applied to a precipitation dataset. The different temporal properties of the dataset motivate the use of some multivariate time interpolation methods like the one proposed in *paper II* and tested in *paper VIII*.

ANN have also been tested, but we arbitrarily consider the same architecture used for the precipitation example. Comparison with other architecture design as well as to extend training time was unfeasible due to the extremely heavy CPU time requirements, a fact related with the size of the testing set as well as the number of input neurons. The set of "optimum" methods in terms of the error measure (RMSE, MAD, p95, etc.) has parameters which were difficult to calculate. This was due to the increased size of the unknown vector **w** and because we used the same limits for the iterations as for the daily precipitation. Future efforts will include the use of high breakdown regression methods (LTS, LMS, etc).

Table 5.14b Preliminary results in m/s for the different imputation methods who uses data from other than the predicted day. The expected value and the median, 75, 85 and 95 percentile of the distribution of the absolute error, and its RMS are presented and compared. In bold the six best results for each estimator.

Algorithm	MAD	Median	75 per cent	85 per cent	95 per cent	RMSE
	m/s	m/s	m/s	m/s	m/s	m/s
bp12	**2.26**	**1.40**	**2.90**	**4.15**	**7.52**	**3.48**
gandin4	**2.22**	**1.34**	**2.81**	**4.12**	**7.66**	**3.47**
gandin6	**2.24**	**1.37**	**2.84**	**4.13**	**7.60**	**3.45**
time_interp	**2.26**	**1.26**	**2.81**	**4.35**	8.28	**3.66**
TIPS (Correlation)	2.52	1.50	3.14	4.68	8.84	3.91
TIPS (Covariance)	2.67	1.66	3.44	4.97	8.92	4.03

5.3.3 Conclusions

The results reported here can be divided in two parts: those originating from early work with the dataset, using a limited number of methods for a single sample of missing values, and those obtained after a Monte Carlo simulation. In the early work we also analyzed the situation of missing values not at random, and concluded that temporal correlation should be taken into account. This result is supported by the results reported from recent studies. The traditional methods of Optimum Interpolation performed well in the Monte Carlo experiment, and taking this into consideration some preliminary conclusions can be drawn:

- in the case of the wind data, the time series structure is very important, even considering only values every three hours.
- the POPS method, based upon minimizing some sort of distance to the center of the cloud, gave poor results.
- the traditional OI considered in *paper VIII* (denoted as gandin20 in table 5.14) is now better than TIPS in terms of the RMSE, but is worse in terms of the robust statistics *Median* of the absolute error and *MAD*. One possible explanation is that the population might be still affected by some remaining outliers, which make results in terms of RMSE very sensitive, but not for the Median or the MAD.
- the weights for the least average, least squares and least 95 percentile methods (which are chosen in order to minimize the MAD, RMSE and p95) might also have been affected by poor choice of the parameters of the solving routine. We have used the same number of iterations, etc. as applied to the precipitation data, and it is clear that some problems arise.
- the ANN performed well, despite that they have been trained using early versions of the dataset (prior and during the correction task) with some outliers included. However, the training time is significantly higher than the time required for daily precipitation, and the observed improvements in the statistics are only of minor significance. Further effort towards better training methods are required. Our primary goal was here to introduce this new technique for the case of meteorological variables.
- use of the historical mean (STAtion aVERAGE) only renders results clearly worse than the other options

5.4 Quality control of raster datasets: the DEM example

5.4.1 Experimental strategy

Again we have applied a Monte Carlo procedure for testing the method with synthetic errors, which were modelled uncorrelated in space or weakly correlated affecting only 1 and 9 pixels in each case. Both cases have been discussed in *paper IV*. The measure of success were mostly the Type I error, and only limited results of the Type II error were reported for the larger errors. Both of them were presented in terms of the effort required for the editing.

In order to gain better insight into the properties of the method, we applied it to a DEM assuming that a second available DEM of higher accuracy is error free. In this case the measure of success was either the RMSE of the errors found or the RMSE of the remaining errors. Neither the Type I nor Type II error as defined in *paper IV* could be considered, because the first is identically zero, while the second decreases linearly with the effort. In both papers a "perfect inspector" hypothesis was assumed for the correction strategies. Here we will present a somewhat more realistic assumption: the inspector will correct a given elevation only once, but he does not have access to the correct elevation. The correction will be done using an interpolated value from the surrounding elevations.

5.4.2 Results for spike-like errors

We will seed the DEM with known errors, and apply the procedure to detect them. We will denote as candidates or guessed errors the set of coordinates (i,j) suggested by any single step of the procedure. The true errors are those candidates that also belong to the known errors set. Fig. 5.15a shows the average Type I error evolution up to 5 per cent depuration effort. The y-axis shows the evolution of the type I error calculated as the number of points misclassified as errors compared with the number of candidates, averaged after 50 replications of the random error set. The x-axis shows the effort, defined as the fraction of the dataset already revised. As before, an effort of 100 per cent implies that all possible points have been checked.

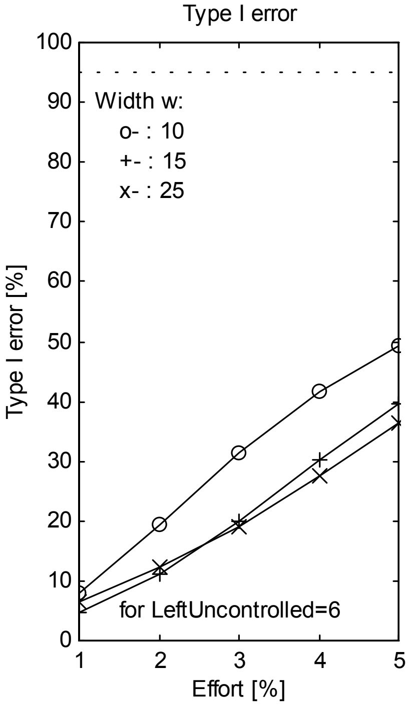

Figure 5.15 Evolution of the Type I error (a) and Type I error (b), as a function of the effort, derived after 50 experiments using spike-like errors. The dotted line in (a) indicated the expected Type I error for a completely random choice. (from paper IV)

We interpolated our results (obtained on a per-step basis) to the prescribed effort values using splines. Each polyline corresponds to different strip width for a fixed number of uncontrolled scores. For w=10, 15 and 25 we left uncontrolled 6 scores. The number 6 can be considered as the limit between the noisy and physically meaningfull scores. In the paper we provide some guidance for its estimation. From this result it is clear that in the first 1 per cent effort the measured Type I error is low, being below 10 per cent for the lower values of w. The dashed horizontal line corresponds to the limit of 95 per cent. Obtaining an error rate over that level is worse than to pick the points at random, since that level is the noise initially seeded into the DEM. The limit was shown being constant, despite the fact that such probability grows slightly as soon as an increasing number of errors has been found. For 2 per cent effort and higher somewhat poorer results may be achieved, but certainly better than chance. But, good results in the type I error is not the whole picture. From a DEM producer's point of view, what is more important is to minimize errors still in the DEM, so the type II error is more representative. Notice that the type I error in fig. 5.15a only counts the successes and the failures, but does not reflect the relative importance of the errors. The absolute value of the error is not considered in the experiment.

We limit our Type II error calculations to errors of absolute size exactly 4m, because the other cases are less prone to be located and its type II error will be unaffected. In fig. 5.15b the evolution of the Type II error is presented.

Notice that the best results are obtained for w=25, but for w=15 the results are similar. The initial Type II error is 1.21 per cent in all cases, so it can be reduced to 0.64 with only 1 per cent effort. The same behavior was noticed for other combinations of w and number of uncontrolled scores.

5.4.3 Results for pyramid-like errors

Since pyramid errors (already presented in 4.3.2 and illustrated in fig. 4.8*)* are spatially autocorrelated, the chance of locating them is lower. The reason for this is that our methods are designed to mainly deal with independent errors in space. We should mention some other details regarding the "accounting" procedure. In figure 4.7 we presented the simplest possibility: every row-wise candidate being also a column-wise candidate is evaluated. This might not represent well the behavior of the operator when faced to spatially correlated errors, because he will analyze also the neighborhood. We will consider as a candidate not only any point which is both a row-wise and column-wise candidate, but also its immediate neighbors. So for every candidate, nine points are checked. However, due to computational simplicity no effort has been made to take into account the overlap of candidates for the same step (i.e. if both a point and its neighbor are selected, there are points that count twice). So the results are somewhat pessimistic in terms of the type I error.

The behavior of the detection procedure (observed in figs. 5.16a and 5.16b) is very similar that observed for the spike-like error shape model, although the numbers are more pessimistic. Up to the first 1 per cent effort, the algorithm renders an acceptable Type I error; however, the error grows slightly with the effort. The analysis of the Type II error should take into account that the initial value in this case is 0.81 per cent, while in fig. 5.15b it was 1.21. This implies that the procedure reduces the Type II error at most 79 per cent with only 1 per cent effort.

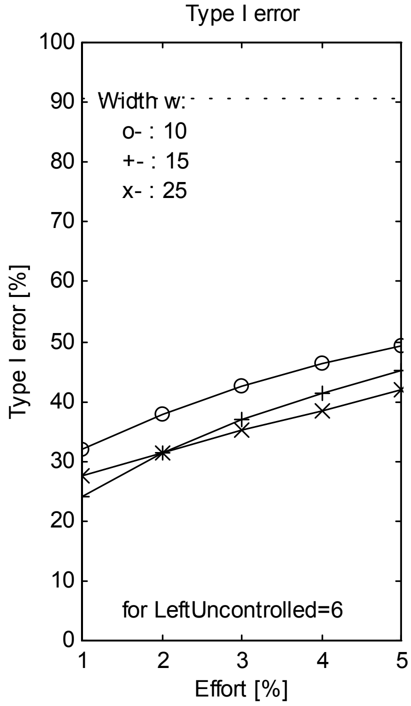

Figure 5.16 Evolution of the Type I error (a) and Type I error (b), as a function of the effort, derived after 50 experiments using pyramid-like errors. The dotted line in (a) indicated the expected Type I error for a completely random choice. (from paper IV)

As expected for the pyramid-shaped error in most of the cases the 1 per cent effort still renders a low type I error. In terms of the type II error, about 21 per cent of the worst errors can certainly be located with only 1 per cent effort of the procedure. This number can rise to 49 per cent with 5 per cent effort. The best results are similar to the ones obtained when dealing with isolated errors, being w=25 the best option. As presented in *paper IV,* the best number of uncontrolled scores is again between 5 and 10, similar to those for the spike-shaped errors.

The conclusion is that the method proves to be effective in identifying a significant amount (up to one third) of the large errors with limited effort. For better Type I results, a smaller w is suggested, while for Type II optimization a somewhat greater w might be the option, irrespective of the shape model assumed. The number of uncontrolled scores is between 5 and 10 in any case. No special pattern of the location of the errors found was noticed. Such aspect have been investigated in the case of real errors as presented below.

5.4.4 Results for real errors

In this case the previously used measures of success are useless. For the simulated case we declare as error everything which differs from the ground truth, and in practice nearly all points of the DEM and its "ground truth" are different. An immediate consequence is that any choice of the locations will succeed in pointing a "true" error, so the Type I error will be identically zero, and the Type II error will decrease linearly with the effort. This preclude to compare results with those presented before, so a different measure of success is required. The most interesting one will should take into account the evolution of the elevation accuracy in terms of the editing effort. Normally, the accuracy of a DEM is not directly known to the user; it can be estimated through sampling in isolated points if more precise measurements are available.

For practical purposes it might be more meaningful to use statistics from the distribution of the errors detected while working with the dataset. For example, its RMS will measure the size of the errors detected by the method up to a given effort. We disregard the RMSE found for each step, because its variability precludes for any simple analysis. Figure 5.17 shows the evolution of the accuracy measured in terms of the RMSE for a strip width of w=8.

The boundaries of the dashed regions at the top and the bottom show the loci for *worst* and *best* possible operation. The former is obtained by considering first the smaller errors, while the latter corresponds to selecting the larger errors first. Under our assumptions both lines should meet at 0 and at 100 per cent. Even though both limits are hardly of practical interest (because it requires knowing the errors in advance) they give a better understanding of the process. Lines with the -⊖- symbol are for the Felicísimo (1994) method, while the others are for different controlled scores using our proposed method. Figure 5.17a show more detail in the low effort region, while figure 5.17b has been extended up to 15 per cent effort. It is clear that the Felicísimo's method outperforms ours in the long run, but at the lower effort region they behave similar. This region is of primary importance for two reasons. Firstly, because most users will not want to go too much further. End users neither have extra data nor the proper tools. They will at most correct the worst errors. DEM producers might go back and make new measurements, but this might become an expensive task if the new values do not differ substantially from the old ones. Secondly, according to Torlegård *et al.* (1986) blunders typically account for less than 3 per cent of the dataset, 0.5 per cent being a median value. Since the methods have been designed for finding gross errors only it is unadvisable to continue the task over such limit.

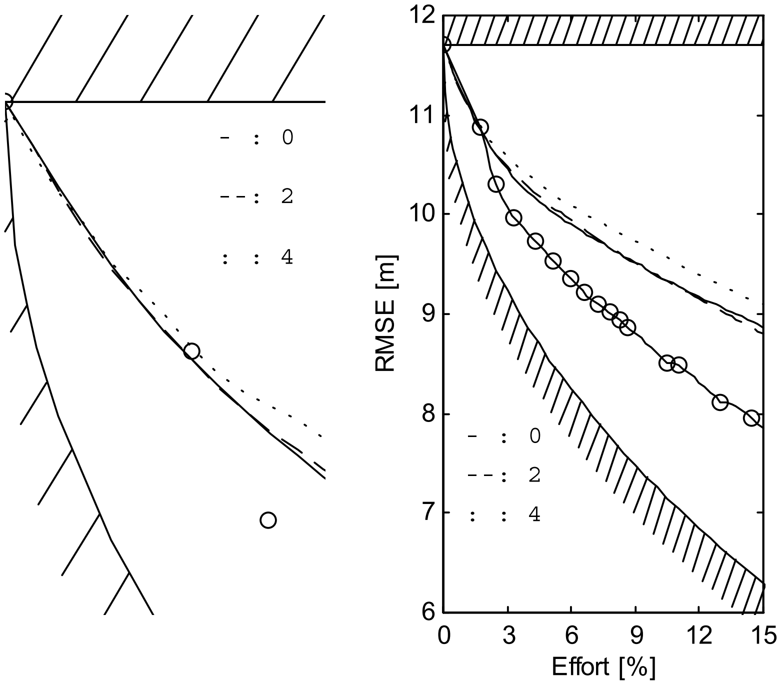

Figure 5.17 Evolution of the accuracy (measured by the RMSE in m) vs. the effort for the methods of Felicísimo (1994) (with the -o- symbol) and the one proposed in paper IV. Results for w=8. Different lines correspond to different number of uncontrolled scores. Left plot shows details of the right one. (from paper V)

It should be noticed that none of the methods shows at 0 per cent effort a slope comparable to the best possible one, which implies that the most important errors are not found in the early stages of the procedure. We also tested some other options for the width parameter w which have not been presented here. The overall result is considered as poor (compared with the conceptual simplicity of Felicísimo's method) and the reason was found to be the "high" spatial correlation of errors. After making the necessary changes (see the paper for further details) to the software described in 4.4.3 we obtained the results presented in figure 5.18.

For this calculation, we subdivided the DEM in regions of width 72 rows, building the strips taking every 9th row within the region. Thus the "strip" width w is again 8. Notice that we skip nearly 10 rows, as suggested by the range of the variogram. The plot of the Felicísimo's method is again included for comparison.

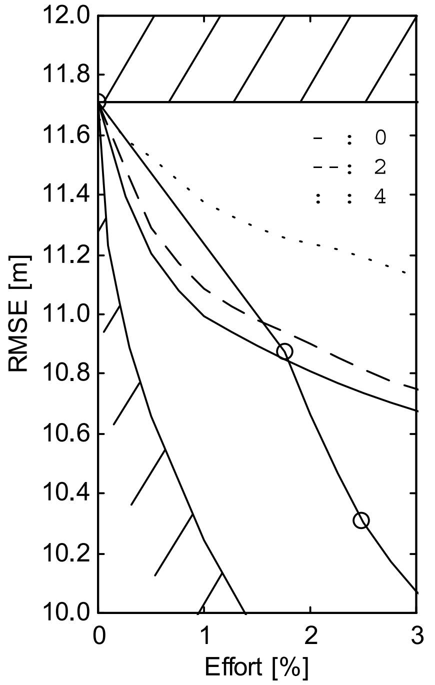

Figure 5.18 Evolution of the accuracy (measured by the RMSE in m) vs. the effort for the methods of Felicísimo (1994) (with the -o- symbol) and the modified one. Results choosing every 9th row. Different lines correspond to different number of uncontrolled scores. Left plot shows details of the right one.(from paper V)

The most striking fact is the difference in the slope at 0 per cent effort, which is markedly closer to the best one. This implies that larger errors are found earlier, leading to a faster decrease of the RMSE. However, once those important errors are removed, the remaining errors are difficult to locate, and the simpler Felicísimo's method is better if the effort exceeds 1.75 per cent.

Neither the end user nor the data producer can draw a plot like fig. 5.18. Instead they can calculate RMS values of the errors already found like those presented in figure 5.19. The x-coordinate is the effort defined as before, while the y-coordinate is the RMSE of the population already corrected. The 0 per cent value is not defined. The figure illustrates the Felicísimo (1994) approach and the modified method of *paper IV*. It is clear that the Felicísimo method finds larger errors in the "long" run (over 1.75 per cent effort) but the modified method is better for lower effort values. Three lines with different number of uncontrolled scores are shown, and it is clear that the one of 0 value is very similar to the one of 2, except very close to the 0.0 per cent effort. The number 2 for the uncontrolled scores were suggested by the empirical rule proposed in *paper IV*.

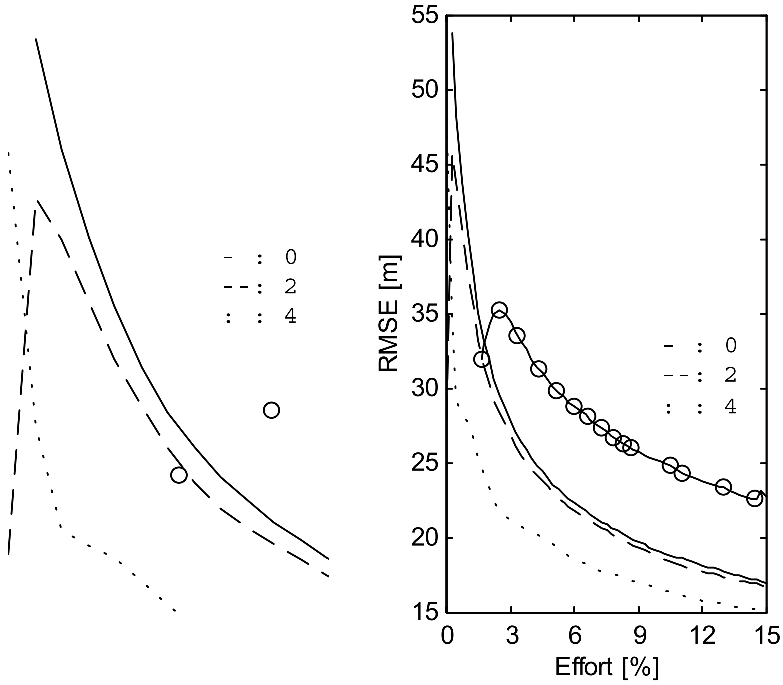

Figure 5.19 Evolution of the RMSE found of the cumulated errors up to a given effort vs. the effort, for the methods of Felicísimo (1994) (with the -o- symbol) and the modified method of paper V. Results choosing every 9th row, resulting in strips of w=8. Left plot shows details (from paper V)

We also analyzed the spatial location of the errors found when a substantial effort has been done. Figure 5.20 shows the places where the Felicísimo's method pointed out the errors up to the 3 per cent effort (in black), and up to 15 per cent effort (in gray). We noticed that most of them are concentrated along significant features of the DEM where slope changes abruptly, namely breaklines. In such points the second order polynomial is not a good approximation of the surface, so differences larger than expected arise. Once early candidate values are corrected, such differences are even more evident. Since we do not allow that any point be corrected twice, its nearest neighbors become candidates. The "clear" image corresponds to early detected points along breaklines, plus its nearest neighbors; there are almost no isolated points. Figure 5.21 shows the pattern for the modified

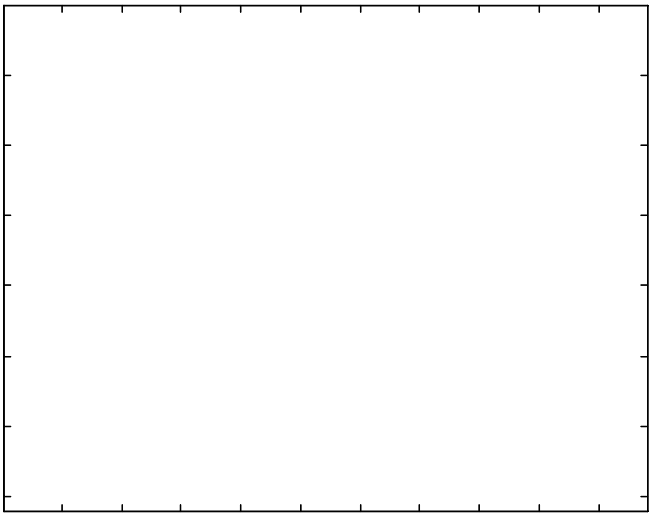

Figure 5.20 Binary map of the errors located up to the 15 per cent effort with the method of Felicísimo (1994). Black areas are for the suggested locations up to the 3 per cent effort; gray ones are obtained after 15 per cent effort

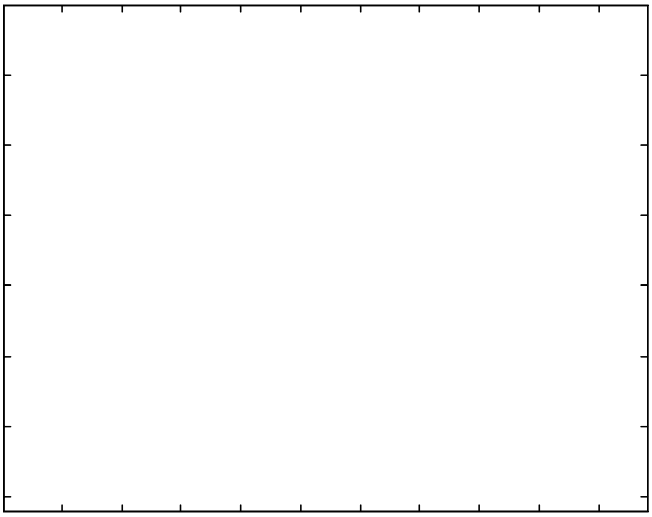

Figure 5.21 Binary map of the errors located up to the 15 per cent effort with the modified method of paper IV. Black areas are for the suggested locations up to the 3 per cent effort; gray ones are obtained after 15 per cent effort

method of *paper IV*. The image looks "noisy" since points are randomly located, and neighborhood is completely ignored.

5.4.5 Results using a "non perfect inspector" assumption

In figures 5.22 and 5.23 we show new results obtained after removing the "perfect inspector" assumption. Under the new hypothesis, the inspector uses an interpolation procedure to correct those values which are believed to be wrong. The procedure uses the elevations in the neighborhood and finds a best fit using second order polynomials (as described for the Felicísimo 1994 method). After an elevation has been corrected, it will not be modified again. If errors are located in adjacent pixels, they will mutually interact masking each other and if the procedure detects one of them, the inspector will produce a modified but not correct interpolated surface. In this case there is no best and worst operation line, because the procedure might in principle introduce new and worse errors in the dataset. Both figures 5.24 and 5.25 were calculated as before using the high accuracy DEM as a reference.

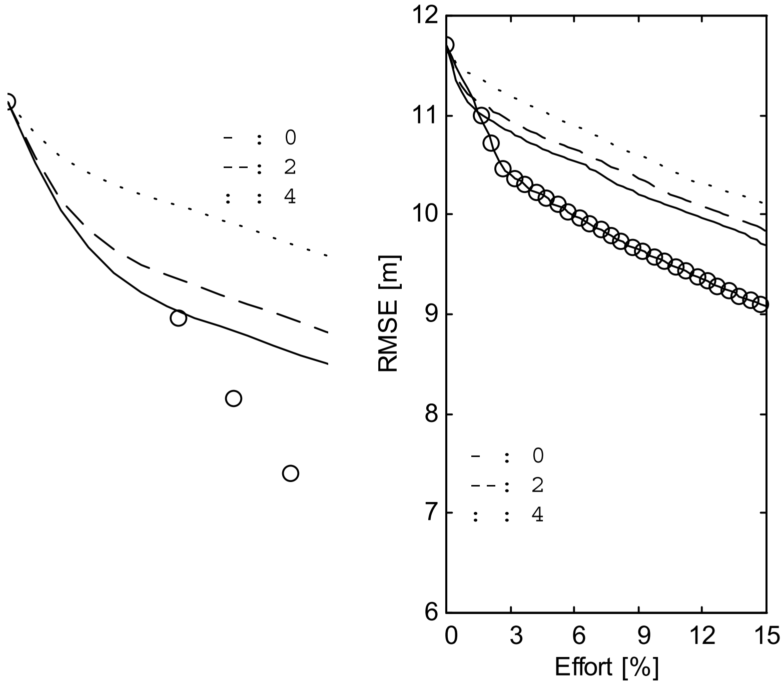

Figure 5.22 Evolution of the accuracy (measured by the RMSE in m) vs. the effort for the method of Felicísimo (1994) (with the -o- symbol) and the modified method. The perfect inspector hypothesis has been removed. Results are shown choosing every 9th row. Left plot shows details of the right one.

The Felicísimo method shows lower performance, mostly due to its preference to pinpoint clusters of elevations. Once an error is imputated using its neighbors it will become a "corrected" point; and if the neighboors are wrong, it will be so. Few improvements in any estimate of accuracy can arise in such case. Our proposed procedure also shows a degraded performance, but to a lesser extent. This can be easily explained analyzing the pattern of candidate locations (fig. 5.21) because our candidates are typically isolated; once imputated (maybe with also wrong values) the next candidate is usually not close to the previous one, and the error correction procedure is not trapped in the neighborhood of few candidates.

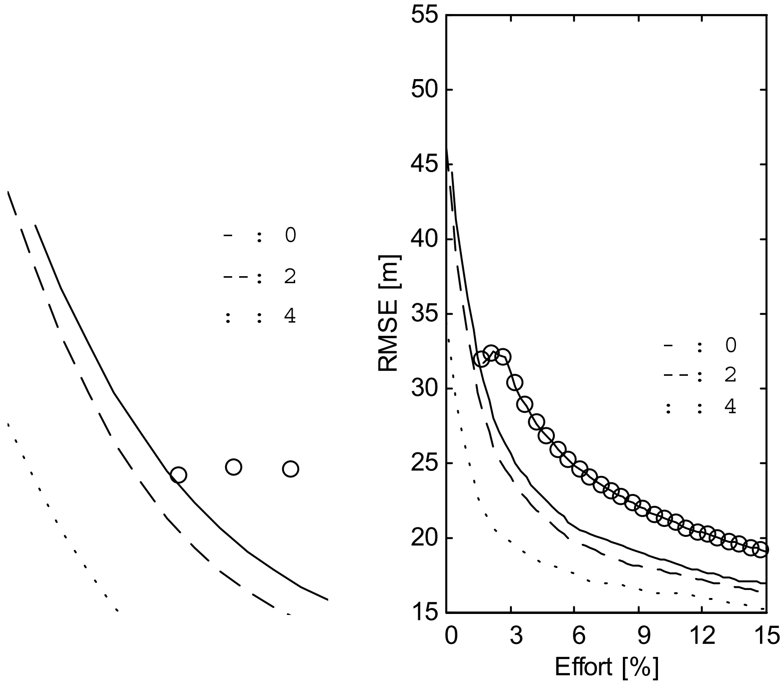

Figure 5.23 Evolution of the RMSE found of the cumulated errors up to a given effort vs. the effort, after removing the hypothesis of the perfect inspector. Plots are for the methods of Felicísimo (1994) (with the -o- symbol) and the modified method of paper V. Results choosing every 9th row. Left plot shows details of the right one

5.4.6 Discussion

The results show that a process can be devised to detect an important part of the larger random errors in a raster dataset. Further actions strongly depend on which type of application the user is involved in. In a production environment, some action can be taken to check the identified isolated values. In photogrammetric measurements these checks can be done before removing the stereo pair from the instrument. The goal here is to improve the accuracy while the effort is less crucial.

The end user is left alone in most cases, because he may not be able to go to the original data sources. Therefore, he will be interested in "evident" errors, i.e. those of relevant size (which are typically few). The results for the synthetic error experiment show that it can be assumed that up to the 1 per cent effort, most candidates are errors. The associated Type I error can be less than 10 per cent, as has been shown for isolated errors, and around 25 per cent for pyramid-like errors with proper choice of the parameters. The Type II error was defined in the first experiment only for errors of absolute size 4 m; it can be reduced 64 per cent (for isolated errors) and 21 per cent (for pyramid-like errors) checking only 1 per cent of the dataset. It is clear that the overall performance decreases as long as the spatial correlation increases.

Every step in the procedure creates a candidate set. Once this set is obtained, any standard procedure can be used to replace the outliers with suitable values. As long as the dataset is being corrected progressively, the risk that a point classified as an error is correct grows, and in practice some caution should be taken.

The test area used in *paper IV* is considered to be a difficult one. Rough terrain, narrow channels, steep hills, and small water areas are typical, all of them may easily mask errors. The DEM itself should not be considered as free of errors, and it has been used "as is". Lack of information about acuraccy is common to most users of this kind of data, so it is believed that this will not limit the range of applications of the ideas presented.

In *paper V* we suggested a modification to the method described in *paper IV*, and carried out a comparative test for both methods and the one suggested by Felicísimo (1994) using real data with known errors. The last method is very simple, but no results using either synthetic or real errors were previously reported. One interesting fact is that this method is parameter free. However, it has been derived under some hypotheses that do not apply to the DEM used in this study. It relies on a low order polynomial interpolator using only nearest neighbors. We think that it will work better in smooth terrain. The use of low order polynomials tends to pinpoint errors which are close to each other, a situation which is more likely to occur with systematic errors. For further work we suggest considering the use of a local Universal Kriging interpolator (Samper and Carrera 1990) using more neighbors. This is in agreement with the findings of Giles and Franklin (1996) who also used a window with 11 by 11 elements. The Kriging approach also allows to model different spatial correlation scales typical for inhomogeneous terrain.

In order to handle the spatial correlation of errors, we have proposed a modification of our method first presented in *paper IV*. We form the strips by subsampling the DEM at each k-th row. From the programming point of view this is a minor change. In real applications, the number k has to be fixed *a priori*. Östman (1987) suggested that k is strongly connected with both the DEM and the acquisition method. The range of the variogram can in practical cases be obtained from errors at control points, or it might be included in the lineage metadata.

Methods described by Felicísimo (1994), *paper IV* and *paper V* have been used in an iterative fashion. Once some errors were removed, all the calculations have been carried out again, and a new candidate set was created. If this set is empty, some parameters are modified automatically (lowering confidence limits, for example) in order to continue the procedure. We continue until 15 per cent of the DEM elevation values has been corrected or confirmed. According to Torlegård *et al.* (1986) gross errors account for less than 3 per cent of the population, so the 15 per cent limit is well within either the systematic (as defined by Thapa and Bossler 1992) or the random error set, provided the first 3 per cent really were gross errors.

It is appropriate here to comment the computer time requirements: the method of Felicísimo (1994) is fairly cheap (of the order of m.n operations, where (m,n) describe the size of the DEM), while both the original and the modified procedure presented involve, for each step, the computation of (m/w).(n/w) covariance matrices of size (w,w), which takes $[(n/w).O(n^2)+(m/w).O(m^2)].O(w^2)$ operations. To calculate the eigenvectors requires $[(n/w)+(m/w)].O(w^2)$ operations, and to project each strip to calculate the scores requires (m+n).w operations. Some other operations are required but depend linearly on m and n. In our example, for a DEM of size m=360, n=216, and for w=8, about 5 minutes per step are required using MATLAB on a SUN Sparc 20. The overall procedure is considered cheap in terms of computer time.

5.4.7 Conclusions

Some methods to locate gross errors in quantitative raster data have been presented, and they were tested in two grid-based DEMs. The DEM of unknown accuracy (analyzed in *paper IV*) has elevations ranging from 0 to 60 m, and it has been used "as is". It has been seeded with artificial errors of low spatial correlation. The DEM with known errors, derived from SPOT data, has elevations ranging from 181 to 1044 m. A more accurate DEM of the same area is available, and it has been considered as the ground truth. The hypothesis of errors uncorrelated in space seems to be wrong at least for this case, as well as the assumption of gaussian distribution for the residuals. This poses serious concern about the usefulness of some previously published algorithms (Felicísimo 1994; *paper IV*) and motivated the *paper V* experiment. The results considering real errors suggest that Felicísimo's method find mostly what is regarded as systematic errors, mainly due to the interpolation algorithm (biquadratic polynomial). The method presented in *paper IV* show similar results in terms of RMS of errors only in early stages of the correction process.

In order to handle the significant spatial correlation observed, a modified version of the second method has been designed and tested with the same dataset. The results were significantly improved and exceeded those of Felicísimo up to a certain level of effort, the effort being defined as the fraction of the DEM elevations corrected or revised. This effort level (1.75 per cent) is of the order of the number of gross errors typically found in DEM; moreover its location pattern looks sparse and random, as opposed to the pattern produced by Felicísimo's method.

The modified method has some free parameters. The most important parameter is an estimation of the correlation lag (the range of the variogram). It can be estimated from a limited number of independent control points. Some authors claim that its value depends on the method for acquisition of the DEM and the DEM itself. In the experiment, the variogram's parameters were supplied by the producer.

We assumed for most of the work that once an error is identified, it can be corrected without error. This is known as the *perfect inspector* hypothesis. We have also tested a more realistic assumption: the inspector will imputate the unlikely value using available tools (usually some interpolation procedure). In the case of using the error detection procedure in a semi-automatic production environment, the method warns the operator about possible errors before the stereopair is unmounted, enabling new measurement. In a fully digital production environment, some correlation thresholds are usually varied to minimize computer time. The method may be used to selectively strengthen the correlation thresholds in suspicious points. In case there is no possibility to verify the errors, e.g. for end users, the algorithm will help to locate the most unlikely values; we have simulated this case under an *imperfect inspector* hypothesis. Unlikely values were imputated using a polynomial fit using nearest data, and the results were qualitatively similar to those obtained using ground truth data for imputation.

5.5 Tabular qualitative data: the national census example

5.5.1 Experimental setup

A Monte Carlo simulation was performed modifying the answers of a subset of the raw data described in Chapter 3. The error model used was very simple. A prescribed number of surveys was chosen at first and then a random number generator chose a fixed number of questions (out of 20) to modify. For each of them, the existing answer was changed to a different value, but still belonging to its feasible set (assuring that they were different from the original one). That was considered a suitable choice for modeling "true" errors. The *total number of contaminated surveys* was fixed as 10, 5, 3 and 1.5 per cent of the subset of 2500 individuals. The figures to be presented correspond to the 3 per cent case, which implies 75 wrong cases out of 2500 surveys.

5.5.2 Results

Figure 5.26 shows a global summary of the behavior of the method. The x-axis is the effort level (already defined in 4.1) while the y-axis represent the fraction of the total errors found. 100 replications of the noisy dataset were used in the Monte Carlo procedure, changing two answers in each survey.

The straight line indicates the locus of the theoretical evolution of the standard (blind) duplicate performance method, i.e. by typing the x per cent of the whole dataset, the same x per cent of the errors were removed (notice that the line goes through the (20 per cent, 20 per cent) point). The dotted line is the best possible operation curve: retype first only those surveys that have errors. It should go through (3 per cent, 100 per cent). Each point of the cloud represents an intermediate situation in a single replication of the Monte Carlo experiment. The figure shows that when retyping 5 per cent of the original data (x-axis) we can locate an amount of the original errors ranging from 25-60 per cent, and when retyping 10 per cent, 40-75 per cent can be located. Further retyping will show a degraded performance, because the "worst" errors have already been located. The limit goal of the procedure will be the (100 per cent, 100 per cent) point, because if all the data are checked we assume that all the errors will be removed. This procedure is intended to be applied for *partial* retyping.

5.5.3 Discussion

Comparing the use of logical edits against the present methodology, some clear differences arise. When the population is updated using mostly the same questions, but with changes in some of them, all related rules should be revised. If a question is ambiguous, the rule can be wrong, while the method proposed in *paper VI* probably will flag the answers as "uncorrelated" and will automatically refuse to control them. The proposed method does not require any expert, since the "rules" (if any) are embedded in the population. Even the dichotomic answers (like marital status, sex, etc.) which are mutually exclusive, are handled gracefully, and need not to be analyzed separately.

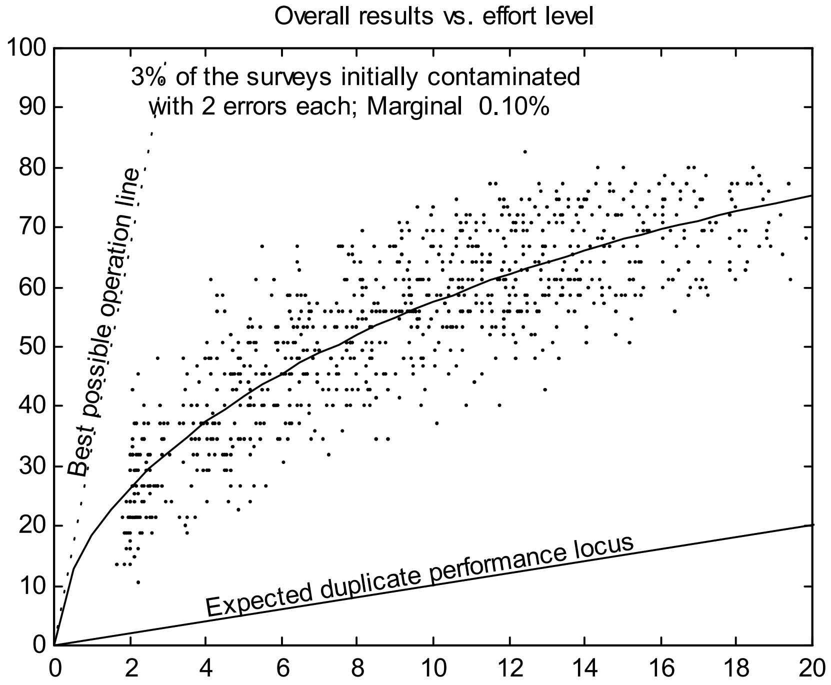

Figure 5.26 Evolution of the remaining errors against the retyping effort for the suggested depuration order and the blind retyping. Plots derived after 100 experiments, modifying 3 per cent of the surveys with 2 errors each (from paper VI)

Since merely retyping is a completely blind methodology, it will locate errors equally well in “unusual” as well as “typical” events, maintaining the variability of the dataset. The proposed method and the logical edits are oriented toward flagging only those errors which make a particular individual “unusual”. Then they will decrease the variability in the dataset.

The application of either logical rules or merely retyping does not require a large population of individuals, while the proposed method implicitly does. Another limitation is that not all the questions can be controlled, either because of almost trivial answers or low correlation with other answers. Moreover, it can not presently handle events with missing values.

The numerical procedure is quite simple. It requires first to transform all categorical answers to a “check box” format, where only ones or zeroes will be admitted as answer. Then, the covariance matrix is constructed and its eigenvectors calculated, and a new table of projections (scores) of the original individuals over the eigenvectors is created. The covariance matrix is no longer positive definite, but is semidefinite. By analyzing the eigenvectors, a critical set of the scores is chosen in order to calculate an outlier region for each. Every event with at least one of its scores lying in those regions should be retyped. The whole procedure can be automated. Once the eigenvectors and the critical set are calculated, they can be applied during the early typing process, allowing for near real time quality control.

The sensitivity of the results to the margin (related with the number of individuals to be retyped in each step) was only weakly significant. The method for selecting the noisy scores to check based on Hawkins (1974) seems feasible, but no further tests have been carried out. As a limiting case, for perfectly uncorrelated answers the procedure is equivalent to looking for answers with low probability, which is also a feasible procedure.

5.5.4 Conclusions

The problem of quality control of categorical data is treated with a method derived from statistical procedures for quantitative tabular data. Two other alternatives have been analyzed; the duplicate performance method and the use of logical edits. The first method is very simple and popular, and requires typing again the same dataset. Its ability to locate errors for a given typing effort is known to be low. The use of logical edits strongly relies on the existence of an expert, which should prepare a set of rules expressed in terms of logical relationships between the answers. When any of them is not met, the survey is flagged as unusual, and either an expert should analyze it or a blind retyping is performed. Here an alternative is proposed in order to carefully reorder what should be retyped.

Some limitations of the proposed procedure are: a large (yet undefined) population is required as well as a minimum number of options for the answers; it cannot handle missing data, and depending on the inherent characteristics of the population, some answers or options for answers are not checked. The users of a method like this are those which are either collecting or using the raw data. We are not providing any tool to check derived statistics (like averages in a region, etc.).

6 Discussion and conclusions

The results presented in section 5 were discussed individually within that section. Here, the results are discussed in the context of the underlying ideas for the thesis, as presented in section 2.

6.1 Where are the outliers that matter?

Most of the effort was directed to adapt methods used to process quantitative tabular data to handle data of different characteristics. The quantitative tabular case has been a subject of interest in multivariate analysis and it will be so in the future. This will assure a continuous flow of new results and methods coming from the statistics area, which hopefully will reinforce our strategy for their integration in GIS.

We analyzed a number of typical examples of tabular quantitative data, taken from the field of meteorology. The datasets consist of measurements taken at regular intervals in time in a prescribed network of weather stations. Many thousands of events are typical in this context, and we limited ourselves to a few thousands. In addition to the widely known, already published methods, we propose new ones, which proved to be successfull. Throughout the work we tried to compare our proposed strategies against other methods. Some early work was methodologically superseded by recent work, and some material has been included in the main body of this thesis for the first time. We bring into consideration a new technique based on Principal Component Analysis (PCA), and another based on Artificial Neural Networks (ANN). The first technique was applied to detect real outliers in a dataset, going back to the original readings existing on paper. The new method as well as the ones existing in the literature were compared through a Monte Carlo experiment. The comparison was made using measures of both the accuracy of the dataset and the relative success in detecting errors, which will provide useful information to end users as well as data producers. Some of the measures have been proposed here for the first time. They measure some kind of distance between the method, the best possible method and the worst possible method. All our previous papers (as well as the main body of the literature) presented the results in terms of the Type I error. We feel that these new measures (valid only for the Monte Carlo procedure) are more informative than either Type I or II errors.

The second case considered was quantitative raster data. We consider the special case of DEM for two main reasons: they are widely used in GIS applications, and there exists a ground truth which can (in principle) be used to correct the errors. We suggested a new approach using a table-oriented method, and compare the results against one of the (few) methods reported in the literature. The experiment was conducted using two DEM for the same area, one of higher accuracy than the other. The first DEM was regarded as the ground truth, while the procedure was applied to the second. The results show that our method results in better accuracy for a given correction effort, provided we correct less than 2 per cent of the points in the DEM. This is approximately the order of the number of gross errors found in practice. Further effort will lead to better performance of the other method. One interesting fact of the new method is that it ignores spatial autocorrelation, but in turn relies on some *parallell correlation*, i.e. correlation between parallell profiles. The pattern of errors located shows that our method found errors in sparse locations. The other method finds mostly systematic errors, which are known to occur where abrupt changes in the slope occur, and are clustered.

The third case considered was tabular qualitative data, like the one collected in population surveys. We have adapted the dataset in order to use any tabular quantitative method. We have again used our PCA based method and compare the results to the traditional duplicate performance method. We performed a Monte Carlo simulation using housing data from Uruguay and found that our method is on average five times better for a given effort than the expected results for the duplicate performance one. From the beginning we discarded methods based on rules, because they do not fit in our framework. However, they are commonly applied in practice, and further comparison should be considered for future work.

6.2 How to fill the gaps?

After an error is detected, analyzed and definitely classified as such, some action need to be taken. The most rational one is to go back and take the measurement again. However, there might be fundamental or practical problems which preclude the use of this solution. The alternative is therefore to fill the gap using the available information. For the case of quantitative raster data the well known technique of kriging might provide an acceptable solution. Since a number of suitable methods exist (some general and some particular for DEM) the raster dataset has not been considered further here.

The qualitative tabular case has also not been addressed as we concentrated on quantitative tabular datasets. We have worked with daily precipitation datasets and with hourly surface wind. Daily rain data have received little attention in the meteorological literature, which mostly considers monthly or even yearly averages only. We have tested a number of different alternatives. The results have been presented in a number of papers, and show that this is a difficult problem. The main reason is that the probability distribution function for the daily precipitation is far from been gaussian, which is a default hypothesis in most of the procedures. Other issues related to the underlying accuracy of the dataset might be raised. It is extremely difficult to decrease the average error below 2 mm/day, while the readings are supposed to be valid to tenths of mm/day (even though the instruments might not be so accurate!). We introduced for the first time non linear methods based on Artificial Neural Networks to this problem, which despite heavy demands on computing power during the training stage should be taken into consideration for future work.

6.3 Concluding remarks

The work presented in this thesis contributes to the development of automated systems for quality control of typical GIS datasets in two aspects:

- By showing that techniques designed for tabular quantitative datasets can be used for at least two other typical data types used in GIS: DEM and qualitative tabular datasets.
- By introducing some new techniques based on PCA and ANN to the problem of quality control of quantitative tabular datasets.

Further work will be required to extend the present results to vector data. The ultimate purpose of this work is to help end users and data producers to improve the accuracy of their datasets, while keeping cost to a minimum.

References

Amhrein, C. and Griffith, D. A., 1987, GIS, Spatial Statistics and Statistical Quality Control *Proceedings of IGIS´87, ASPRS/ACSM, Falls Church, VA.*

Anonymous, 1994, ESRI tops the charts again. *ESRI ARC News*, **16,** 3, 35-35

Bethel, J. S.; Mikhail, E. M., 1984, Terrain surface approximation and on-line quality assessment. International Archives of Photogrammetry and Remote Sensing. Commission III, V25, A3a, 23-32

Borovkov, A. 1987, Statistique Mathématique. Editions MIR (Moscou) 600 pp.

Bruce, C.; Eischeid, J. K.; Karl, T.R. and Díaz, H.F., 1995, The quality control of long-term climatological data using objective data analysis. *Preprints of the AMS Ninth. Conf. on Applied Meteorology, Dallas, TX., Jan 15-20.*

Buttenfield, B. P. (editor), 1993, Mapping data quality. *Cartographica*, **30**, 2-3, 1-45

Cybenko, G., 1989, Approximation by Superpositions of a Sigmoidal Function *Math. Control Signals Systems,* **2**, 303-314

Dahlquist, G.; Björck, Å. and Anderson, N., 1974, Numerical Methods, Prentice Hall, Englewood Cliffs, N. J., 572 p.

Davies, L. and Gather, U., 1993, The identification of multiple outliers. *Journal of the American Statistical Association,* **88**, 423, 782-801

Day, T. and Muller, J.P., 1988, Quality assessment of Digital Elevation Models produced by automatic stereo matchers from SPOT image pairs. In *Proceedings of the 16th. International Congress of the International Society for Photogrammetry and Remote Sensing, Kyoto*. Commission III, 148-159.

Demuth, H. and Beale, M., 1994, Neural Network User's guide (Toolbox for MATLAB) The MathWorks, Inc. 226 pp., http://www.mathworks.com

Eskridge, R. E.; Alduchov, O.. A.; Chernykh, I. V; Panmao, Z.; Polansky, A. C. and Doty, S. R., 1995, A comprehensive Aerological Reference Data Set (CARDS); Rough and Systematic Errors. *Bulletin of the American Meteorological Society*, **76**, 10, 1759-1775.

Felicísimo, A., 1994, Parametric statistical method for error detection in digital elevation models. *ISPRS J. of Photogrammetry and Remote Sensing*, **49**, 4, 29-33.

Fellegi, P. and Holt, D., 1976, A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, **71**, 353, 17-35

Gandin, L. M., 1965, Objective analysis of Meteorological Fields. *Israel Program for Scientific Translations*, 242 pp.

Gandin, L., 1988, Complex Quality Control of Meteorological Observations. *Monthly Weather Rev*iew, **116**, 1137-1156

Giles, P.T. and Franklin, S.E., 1996, Comparison of derivative topographic surfaces of a DEM generated from stereoscopic SPOT images with field measurements *Photogrammetric Engineering & Remote Sensing*, **62**, 10, 1165-1171

Goodchild, M. F. and Gopal, S., 1989, Learning to live with errors in Spatial Databases *Accuracy of Spatial Databases Preface* Taylor and Francis, London, 290 p.

Goodchild, M. F., and Hunter, G. J., 1997, A simple positional accuracy measure for linear features To appear in *International Journal of Geographical Information Science,* **11**, 3

Goodchild, M. F.; Guoqing, S. and Shiren, Y., 1992, Development and test of an error model for categorical data *International Journal of Geographical Information Systems* **6**, 2, 87-104

Griewank, A., Juedes, D. and Utke, J., 1996, ADOL-C A package for Automatic Diferentiation of Algorithms written in C/C++. *ftp://info.mcs.anl.gov/pub/ADOLC/ADOLC_BETA/adolc.ps.Z; also* FORTRAN 90 in *adolf.ps.Z.*

Gutiérrez, C., 1996, Análise de uma Metodología para o Recheio de "Missing Values" numa Base de Dados de Chuva, Baseada na Pseudo-Distancia de Kulback-Leibler *Proceedings of the IX Congresso Brasileiro de Meteorologia, Campos do Jordao (in portuguese)*, 253-257

Hadi, A. S., 1994, A Modification of a Method for the detection of Outliers in Multivariate Samples. *J. Royal Statist. Soc. B* **56**, 2, 393-396

Hadi, A.. S., 1992, Identifying Multiple Outliers in Multivariate Data. *J. Royal Statist. Soc. B* **54**, 3, 761-771

Hadi, A.. S., 1997, Personal communication

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

Hawkins, D. M., 1993, The feasible set algorithm for least median of squares regression. *Computational Statistics & Data Analysis*, **16,** 81-101.

Hawkins, D. M., 1994a, The feasible set algorithm for least trimmed squares regression. *Computational Statistics & Data Analysis*, **17,** 185-196.

Hawkins, D. M., 1994b, The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics & Data Analysis*, **17,** 197-210.

Hawkins, D. M., 1997, Personal communication

Hunter, G. and Goodchild, M., 1996, Comunicating uncertainty in spatial databases. *Transactions in GIS,* **1,** 1, 13-24

Hunter, G. J. and Beard, K., 1992, Understanding error in Spatial Databases *The Australian Surveyor,* **37**, 2, 108-119

John, S. A., 1993, Data integration in a GIS - The question of data quality. *ASLIB Proceedings*, **45**, 4, 109-119.

Johnson, G. T., 1982, Climatological Interpolation Functions for Mesoscale Wind Fields. *Journal of Applied Meteorology*, **21**, 8, 1130-1136.

Johnsson, K., 1994, Integrated digital analysis of regions in remotely sensed imagery and thematic data layers. *Ph.D. Dissertation TRITA-FMI Report 1994:2,* Dept. of Geodesy and Photogrammetry, Royal Institute of Technology, Stockholm, Sweden

Kanevsky, M.; Arutyunyan, R.; Bolshov, L.; Demyanov, V. and Maignan, M., 1996, Artificial neural networks and spatial estimations of Chernobyl fallout. *Geoinformatics*, **7**, 1-2.,5-11.

Keefer, B. J.; Smith, J. L. and Gregoire, T. G., 1988, Simulating manual digitizing error with statistical models. *Proceedings of GIS/LIS '88 Falls Church, VA: ASPRS/ACSM*, **2**, 475-483

Koroliuk, V. S., Manual de Teoría de Proabilidades y Estadística Matemática. Ed. MIR, Moscow, 580 pp. (in spanish)

Lebart, L.; Morineau, A.; Tabard, N., 1977, Techniques de la description statistique: Methodes et logiciels pour l'analyse des grands tableaux. Ed. Dunod, Paris, 344 pp.

Little, R. J. A., 1988, A Test of Missing Completely at Random for Multivariate Data with Missing Values *Journal of the American Statistical Association,* **83**, 404, 1198-1202

López, C. and others., 1997, Informe final del proyecto BID/CONICYT 51/94 (in spanish, with english abstract)

Maronna, R. A. and Yohai, V. J., 1995, The Behavior of the Stahel-Donoho Robust Multivariate Estimator *Journal of the American Statistical Association*, **90**, 429, 330-341

Maronna, R., 1976, Robust M-estimators of Multivariate location and Scatter, *The Annals of Statistics,* **4**, 1, 51-67

Moore, R. E., 1966, Interval Analysis, Prentice Hall, Englewood Cliffs, N. J.

Nebert, D., 1995, Status of the National Geospatial Data Clearinghouse on the Internet. In *Proc. of the 15th. Annual ESRI User Conf., May 1995*. Also http://www.fgdc.gov/clearinghouse/pubs/esri95/p196.html

Nebert, D., 1996, Supporting Search for Spatial Data on the Internet: What it means to be a Clearinghouse node. *In Proc. of the 16th Annual ESRI User Conf.* Also http://www.fgdc.gov/clearinghouse/pubs/esri96/revised.html

Nychka, D. and O'Connell, M., 1996, Neural Networks in Applied Statistics: Discussion *Technometrics,* **38**, 3, 218-220

Openshaw, S., 1989, Learning to live with errors in Spatial Databases *Accuracy of Spatial Databases* (M. Goodchild and S. Gopal, editors) Taylor and Francis, London, 263-276.

Östman, A., 1987, Quality control of Photogrammetrically sampled Digital Elevation Models. *Photogrammetric Record*, **12**(69), 333-341.

Reek, T.; Doty, S. R. and Owen, T. W., 1992. A Deterministic Approach to the Validation of Historical Daily Temperature and Precipitation Data from the Cooperative Network. *Bull. Amer. Met. Soc.*, **73**, 6, 753-762

Richman, M. B., 1986, Review article: Rotation of principal components. *Journal of Climatology*, **6**, 293-335.

Rocke, D. M. and Woodruff, D. L., 1996, Identification of outliers in Multivariate Data *Journal of the American Statistical Association*, **91**, 435, 1047-1061

Rocke, D. M., 1996, Robustness properties of S-estimators of Multivariate location and shape in High dimension, *The Annals of Statistics,* **24**, 3, 1327-1345

Rocke, D. M., 1997, Personal communication

Rousseeuw, P. J. and Leroy, A., 1987, Robust Regression and Outlier Detection, New York: John Wiley

Rousseeuw, P. J. and Van Zomeren, B.C., 1990, Unmasking Multivariate Outliers and Leverage Points *Journal of the American Statistical Association,* **85**, 411, 633-639

Rousseeuw, P. J., 1984, Least Median of Squares Regression *Journal of the American Statistical Association,* **79**, 388, 871-880

Rousseeuw, P. J., 1991, A Diagnostic Plot for Regression Outliers and Leverage Points *Comput. Statistics & Data Analysis,* **11**, 127-129

Rumelhart, D. E.; Hinton, G. E. and Williams, R. J., 1986, Learning representations by Back-Propagating errors *Nature,* **323**, 533-536

Samper, F.J. and Carrera, J., 1990, Geoestadística: aplicaciones a la hidrología subterránea. ISBN 84-404-6045-7 480 pp (in spanish)

Samper, J. and Neuman, S. P., 1989, Estimation of Spatial Covariance Structures by Adjoint State Maximum Likelihood Cross Validation. 1. Theory. *Water Resour. Res*. **25**, 3, 351-362

Stern, H., 1996, Neural Networks in Applied Statistics *Technometrics,* **38**, 3, 205-220 (with discussions)

Strayhorn, J. M., 1990, Estimating the errors remaining in a data set: techniques for quality control, *The American Statistician*, **44**, 1, 14-18

Thapa, K. and Bossler, J., 1992, Review article: Accuracy of spatial data used in Geographic Information Systems. *Photogrammetric Engineering & Remote Sensing*, **58**, 6, 835-841

Torlegård, K; Östman, A. and Lindgren, R., 1986, A comparative test of photogrammetrically sampled digital elevation models. *Photogrammetria*, **41**(1), 1-16

Warner, B. and Misra, M., 1996, Understanding Neural Networks as Statistical Tools *The American Statistician,* **50**, 4, 284-293

Appendix 1

López, C., González, J. F. and Curbelo, R., 1994, "Principal component analysis of pluviometric data a) Application to outlier detection" *English translation of the paper* "Análisis por componentes principales de datos pluviométricos. a) Aplicación a la detección de datos anómalos" *Estadística (Journal of the Inter-American Statistical Institute) 46, 146,-147, pp. 25-54*

Principal component analysis of pluviometric data
a) Application to outlier detection

English translation of the paper "Análisis por componentes principales de datos pluviométricos. a) Aplicación a la detección de datos anómalos" *Estadística (Journal of the Inter-American Statistical Institute) 46, 146,-147, pp. 25-54.*

Carlos López [1], Elizabeth González[2]
and Jorge Goyret[1]

Abstract

The techniques used for the treatment of a pluviometric data bank during the development and calibration phase of a flow-rain, flow hydrological model are presented.

The calibration phase of this type of models is considerably affected by errors (*outliers*) in the calibration set. Thus it is mandatory to either correct or eliminate those records. We applied a variety of methods for this dataset. Among them, the Principal Component Analysis (PCA) gave the best results.

The developed methodology allows real time quality control of newly acquired data with minimum computer resources requirements, which makes feasible its application in standard equipment. For the present paper, we have defined as errors only those records which differ from the value written down on paper by the observer.

However, it is believed that the PCA is able to detect also other random errors from the observer and even some type of systematic ones, which are still in the investigation phase.

1. Introduction

1.1 Sketch of the problem

In all datasets it exists at least two sources for errors: those intrinsic to the measurement operation and those generated either while keying in or during later process of the information. Both types of errors might have an important effect depending on the particular problem. According to Husain, 1989, "... the failure of many projects of considerable budget can be attributed at least in part, to the imprecision of the hydrologic information available...". In the hydrological model case, the errors propagate themselves in time, and depending on the particular characteristics of the catchment area, its effect might be considerable after significant time lags.

In the daily operation of those models, it is fairly simple for the user to notice significant outliers, because a direct evaluation can be done the day after.

[1] Centro de Cálculo, Facultad de Ingeniería, CC 30, Montevideo, Uruguay

[2] Instituto de Mecánica de los Fluidos e Ingeniería Ambiental, Facultad de Ingeniería, CC 30, Montevideo, Uruguay

In turn, during the calibration stage of the model, many empirical parameters must be fixed by analyzing thousands of values of measured vs. calculated flow; this comparison can only be made by analyzing global statistics like the standard deviation, etc.

Such fact mix those events obviously erroneous as well as other more subtle ones, which might lead to significant (and uncontrolled) bias in the parameters. For depuration purposes, it has been assumed that values written on paper by the observer are error free, so we try to detect only typing errors. However it will be clear that the method can be easily extended for handling both random and some systematic errors, due to inappropriate sheltering of the instrument the latter and careless operation the former.

The present work should be considered as a natural extension of the task performed during the calibration phase of an hydrologic model of flow-rain, flow type for the Río Negro catchment area. For further details please refer to Silveira *et al.* (1992a y 1992b).

1.2 Methodological background

Regarding outlier detection procedures, the single national registered reference is due to the guidelines produced by the Climatology Department of the Uruguayan National Meteorological Bureau (DNM, 1988). Those specifically related with rain data are very wide and they are mostly connected with the specification of an admissible range.

At an international level, some comprehensive meteorological work has been published (Sevruk, 1982) in order to correct typical systematic errors in each station. In order to do this, they also require values of the surface wind velocity, rain rate, temperature and humidity of the air, etc.

Regarding random errors, the trend is to compare the direct measurements with a model of the phenomena (see for example, Francis, 1986; Hollingsworth *et al.,* 1986). The latter pointed out that for the case of the surface wind, the differences between observations and predictions follow approximately a gaussian distribution. In that case it is relatively simple to detect outlier values in order to analyze them carefully. An important disadvantage of this approach is the considerable amount of information required, as well as the important computer resources involved.

If we disregard (or simply it is unknown) the physical relationship between the variables, the strictly statistical procedures have to be considered. Barnett *et al*., 1984 reviewed and summarizes many techniques which might be of use in this problem. For the multivariate analysis of data he distinguishes two main methodological trends, depending on the fact wether the probability density function is assumed or not.

The first group techniques are named Discordancy Tests; they require an estimation of the parameter of the distribution. There is also some work which assumes that the theoretical distribution has one shape, and the sample another, as proposed for example by O'Hagan, 1990. He applied the idea for an example involving both a Gaussian and a Student's t distribution. Some rules might help in those cases to highlight outlying values. Our case of daily rain rate do not fit readily under such hypothesis, as follows from a simple analysis of its distribution.

The second group identified by Barnett is named as Informal Methods. They neglect the formal aspects of the probability density function, and attempt in turn to exploit certain properties of the distribution. This group includes graphic methods which look for points far from the data cloud; correlation methods, like those described by Gnanadesikan *et al*., 1972; use of representative generalized distances, techniques usually connected with cluster analysis (see Fernau *et al.,* 1990) and Principal Component Analysis (PCA) among others, etc.

A specific reference related to PCA is the one due to Hawkins, 1974. The author compares four statistics designed to highlight outliers. Hawkins assumed that each observation belongs to a gaussian distribution, an hypothesis which do not hold for the rain; however the concepts that can be derived are similar to the one considered here as well as the results obtained working with coal samples.

2. PCA in brief

PCA is a widely applied multivariate technique (see Richman, 1986 as a general review; Pio *et al*., 1989 for air pollution; White, 1991 for rain, etc.). It might transform one set of correlated measurements into new series of uncorrelated readings, which in turn let consider each one as an independent variable.

Moreover, the new variables minimize the remaining RMS. which might be helpful to distinguish the physics from the noise. In this work we did not attempt to rotate the obtained components, as suggested by Richman, 1986; White, 1991 among others, a process which is supposed to improve the interpretability of components more related with the physics.

2.1 Theoretical aspects

Hereinafter we will denote as $p_i(\tau_k)$ the precipitation value for time τ_k (k=1..r) at station i (i=1..n). The temporal mean at station i will be denoted with an overbar, $\overline{p}_i$.

Given a set of rain readings for a given time $p_i(\tau_k)$ they can be represented together by a vector $\mathrm{P}_{(n,1)}(\tau_k)$ which belongs to the R^n space (fig. 1). Each k-th point of the cloud corresponds to a date τ_k. The origin of coordinates is taken at the baricenter of the cloud, with components $\overline{p}_i$ which will be denoted as P_M.

It is possible to show that it exists a direction $\vec{e}_1$ (unique in the general case) which minimizes the sum of squares S_1

$$S_1 = \sum_{k=1}^{r} \overline{M_k H_k}^2 \tag{1}$$

as sketched in fig. 1. The direction $\vec{e}_1$ does not depend on time τ_k. It will be denoted as $a_1(\tau_k)$ the projection OH_k, which is also named score in the literature. Each term in S_1 can be interpreted as the L^2 norm of the vector

$$\mathrm{P}(\tau_k) - \mathrm{P}_\mathrm{M} - a_1(\tau_k).\vec{e}_1 \tag{2}$$

This expression shows that for any time τ_k the data vector is explained as the sum of a constant vector plus a multiple of a constant vector. The statistic S_1/r can be interpreted as the unexplained variance by an approximation by a single term.

Similarly a vector $\vec{e}_2$ can be found in order to minimize the remaining variance, so

$$S_2 = \sum_{k=1}^{r} \left| P(\tau_k) - P_M - a_1(\tau_k).\vec{e}_1 - a_2(\tau_k).\vec{e}_2 \right|^2 \tag{3}$$

being $a_2(\tau_k)$ the projection over the direction $\vec{e}_2$ of the vector OM$_k$. Even from geometric arguments it can be shown that $\vec{e}_1.\vec{e}_2 = 0$.

We can apply the procedure up to S$_n$. Lebart *et al.*(1977) demonstrates that $\vec{e}_i$ are eigenvectors of the covariance matrix, defined as

$$C = \left\{ c_{ij} : c_{ij} = \sum_k \left(p_i(\tau_k) - \bar{p}_i\right).\left(p_j(\tau_k) - \bar{p}_j\right) \right\} \tag{4}$$

and that the eigenvalues λ_i are directly related with the sum S$_i$. It can be shown that the scores time series $a_i(\tau)$ and $a_j(\tau), i \neq j$, have null crosscorrelation. If we denote as D the diagonal matrix with the eigenvalues λ_i in the diagonal, and E the matrix holding the eigenvectors $\vec{e}_i$ as columns, we can prove:

$$C = E.D.E^T \tag{5}$$

In what follows we will use the term *principal components* to refer to the eigenvectors $\vec{e}_i$, and as scores the time series of the associated projections $a_i(\tau)$. It should be noticed that the index i is not related with a pluviometric station.

Summing up, it exists a lineal transformation which relates the observed time series $p_i(\tau), i = 1..n$, with the scores $a_i(\tau)$ which can be written in matrix form as

$$P(\tau) = P_M + E.A(\tau) \tag{6}$$

being P_M the vector holding the mean precipitation of the period, and $A(\tau)$ a vector holding the scores.

$$P(\tau) = \begin{bmatrix} p_1(\tau) \\ \vdots \\ \vdots \\ \vdots \\ p_n(\tau) \end{bmatrix}; P_M = \begin{bmatrix} \bar{p}_1 \\ \vdots \\ \vdots \\ \vdots \\ \bar{p}_n \end{bmatrix}; A(\tau) = \begin{bmatrix} a_1(\tau) \\ \vdots \\ \vdots \\ \vdots \\ a_n(\tau) \end{bmatrix}; E = \begin{bmatrix} \vdots & & & \vdots \\ \vdots & & & \vdots \\ \vec{e}_1 & \vec{e}_n & \cdots & \vec{e}_{n-1} & \vec{e}_n \\ \vdots & & & \vdots \\ \vdots & & & \vdots \end{bmatrix} \tag{7}$$

Except in pathological cases, matrix E is not singular, thus once the rain measurements $p_i(\tau), i = 1..n$ are given, it is possible to obtain the scores $A(\tau)$ by applying the expression:

$$A(\tau) = E^{-1}.\left(P(\tau) - P_M\right) \tag{8}$$

It will be useful later to show that Eq. 6 can be rewritten as

$$P(\tau) = P_M + \sum_{i=1}^{i=n} a_i(\tau).\vec{e}_i \tag{9}$$

2.2 The need for a progressive depuration

The eigenvectors $\vec{e}_{\mathrm{i}}$ (denoted also as *patterns*) are calculated using an available cloud of data points. It might exist a small number of unlikely values (outliers) which might affect to some extent the patterns themselves. In Silveira *et al*., 1991 it has been shown that even in a population of r=4000 points, only two outliers might significantly affect the patterns. This fact makes mandatory that a recursive depuration effort strategy is to be adopted. On early stages we will look only for those more evident values. As it will be shown, the process can go later to look for more subtle cases.

3. Application of the technique to a particular case: the Río Tacuarembó catchment area

3.1 General characteristics of the study area

Despite the work considered a substantially greater area, we restrict ourselves for this analysis to the Río Tacuarembó catchment area, of 20.000 km^2, located at 32° S, 55°W at 400 km of the Atlantic Ocean. The area can be characterized by a smooth orography, with heights lower than 500 m, few valleys and lakes. The monthly mean for the rain is within 74 and 120 mm/month. The study period is nearly of 15 years, from Jan 1$^{st.}$ 1975 to Dec 2$^{nd.}$ 1989, value clearly over the threshold suggested by Hawkins, 1974.

3.2 A brief description of the compared methods

a) Outlying values of the univariate series

This method is fairly simple, and requires the calculation of a "feasible" range for the values recorded in each station: whenever any record is outside it, it is pointed out as a candidate to be in error. In the given dataset it is usual to mistype records taken in mm/day as taken in tenths of mm/day. This values could be found only if the mistyped record is over 100 mm/day, but the procedure is impractical for other cases.

For the daily rain example this method can detect only events clearly outlying by excess, but on the other hand it is impossible to suggest a zero value reading as an error, because over 80% of the population is exactly zero.

b) Discrepancy of the Thiessen's spatial mean series

The first stage requires that the mean average of the rain is calculated by the Thiessen method (Jácome Sarmento *et al.,* 1990) using different subsets of stations taken from the n available for each day. Thus different time series arise, and when compared if they differ "too much" the particular day is checked.

The results obtained (not presented here) let say that this method gives a much powerful test that the one before; true errors exist in nearly 30% of the selected dates (Silveira *et al*., 1991). Moreover, the errors themselves need not to be outlying.

c) Outlying values of the multivariate series

For the Río Tacuarembó dataset, typically two out three days have some missing value. Then, we must distinguish two situations for each time τ:

c.1) All n stations have readings
c.2) Some values are missing

In the first case, it is possible to calculate the n scores $a_i(\tau)$. If for some i, $a_i(\tau)$ is not within the i-th specified range, all n records used for calculate the scores should be checked. The specified ranges were determined by analyzing the probability density function of the scores for the whole period.

In the second case, an imputation procedure is required. It might be nearest neighbor or any other. Using the same dataset López *et al.,* 1994 analyzed the performance of four methods for missing value imputation are compared, being the most efficient the Penalty of Principal Components, so we apply it here.

Once imputated the missing values, we are in the position to apply the criteria of c.1) by checking each of the scores. However, both here and at Silveira *et al.,* 1991 we relaxed the criteria, and the date was checked if any of the imputated records is negative or bigger than 100 mm/day. For further details, please see López *et al.,* 1994.

In figures 3 and 4 the typical probability density function (pdf) for both the weakest and strongest scores are shown. For the range determination, we restrict ourselves to symmetric ones with a single parameter α_i. For each i, α_i is selected in order to make valid over 96% of the events. If the pdf is nearly symmetric (as for example patterns 2, 3, ... 17, see fig. 3 and 4) this rule implies to reject approximately 2% of each tail of the distribution. For heavily skewed distributions (pattern 1, fig. 3) we reject only from one side of the pdf.

4. Results

4.1 With simulated errors

In order to test the ability of the method for this problem, we select a subset of n = 13 carefully revised stations which have less than 5% of missing records for the period of r = 5450 days (nearly 15 years).

We selected at random a set of 2832 triplets of station-date-value which is around 4% of the population. The wrong values for rain were generated by a mechanism which attempts to replicate the pdf of the real data. In fig. 5 we show the distribution for positive values.

We applied the suggested method in order to detect the artificial errors. In tables I and II we presented the total number of error detected discriminated by step. Between the first and second column, the difference is in the recalculation of the limits α_i. The detected errors in the first columns were ignored in order to calculate the new α_i, but they are expected to be detected in the second sweep. Neither the covariance matrix nor the eigenvectors were recalculated.

	First sweep $\frac{\text{total detected}}{\text{total revised}}.10^3$	*Second sweep* $\frac{\text{total detected}}{\text{total revised}}.10^3$
First depuration	$\frac{360+\mathbf{211}}{7644+\mathbf{6318}}.10^3=41$	$\frac{2065+\mathbf{215}}{54067+\mathbf{6435}}.10^3=38$
Second depuration	$\frac{151+\mathbf{35}}{5798+\mathbf{5863}}.10^3=16$	$\frac{1784+\mathbf{40}}{50924+\mathbf{5837}}.10^3=32$

Third depuration	$\frac{68+\mathbf{36}}{4966+\mathbf{7514}}.10^3=8$	$\frac{276+\mathbf{39}}{9555+\mathbf{7397}}.10^3=19$

Table I: Evolution of the depuration process in relation with the data to be checked. Terms in the table follow the schema (A+B)/(C+D).10^3, being A: wrong values detected in full days; B: wrong values detected in incomplete days (**in bold**); C: number of records revised in complete days and D: number of records revised in incomplete days (**in bold**).

Another possibility is detect-correct-recalculate. The results are presented in the first column. For the second depuration, we eliminate the outlying values detected and both the covariance and its eigenvectors are recalculated.

We show in bold the results for days with missing values. In table I, we express the results in relation to the *number of revised values checked against paper*. In table II, we present the results in relation with the total *number of errors yet in the population*.

The results show that it is more convenient to change the limits α_i rather than recalculate the eigenvectors. Thus for two sweeps it can be found 81% of the wrong values, which affects 49% of the revised days.

If we want to recalculate the pattern as soon as we detect the first 571 errors, in the second depuration we found only 186 errors, which account only for 21% of the days to check.

Such behavior was not observed while working with the raw data: even very few errors affected significantly the patterns, requiring in turn a couple of iterations in order to stabilize them.

	First sweep $\frac{\text{total detected}}{\text{total not yet found}}$	*Second sweep* $\frac{\text{total detected}}{\text{total not yet found}}$
First depuration	$\frac{360+\mathbf{211}}{2832}=\frac{571}{2832}=0.20$	$\frac{2065+\mathbf{215}}{2832}=\frac{2280}{2832}=0.81$
Second depuration	$\frac{151+\mathbf{35}}{2832-571}=\frac{186}{2261}=0.08$	$\frac{1784+\mathbf{40}}{2261}=\frac{1824}{2261}=0.81$
Third depuration	$\frac{68+\mathbf{36}}{2261-186}=\frac{104}{2075}=0.05$	$\frac{276+\mathbf{39}}{2075}=\frac{315}{2075}=0.15$

TableII: Evolution of the depuration process in relation with the remaining errors. Terms in the table follow the schema (A+B)/C, being A: wrong values detected in full days; B: wrong values detected in incomplete days (**in bold**) and C: wrong values in the database yet to be found.

4.2 Over real errors

In a real situation a table like Table II cannot be created. It is required also a criteria to stop the procedure: we decided to stop as soon as no new errors (*true* errors) were found. We define as *true* error all those cases which the number in the files do not coincide with the one written on paper.

In early stages we worked for full days (with no missing values) over a set of 21 stations. Two phases could also be distinguished.

In the first one, after performing the PCA calculations, we removed the worse errors. They were identified because even not significantly affecting the mean vector, the first and second patterns were completely distorted (see Silveira *et al.*, 1991 for details). This stage corresponds with rows 1, 2 and 3 of Table III.

In the second phase we selected those days which scores $a_i(\tau_k)$ exceed the allowable value. The measurements for such day were checked against paper, and corrected if any discrepancy exist. Then we recalculate the Principal Components and the process start again.

For each score a_i the limits were estimated either as three times the standard deviation, or were simply ignored. Despite the criteria of the *three times* is a well known boundary valid for the gaussian distribution, the method do not requires neither imply it.

During the task it has been observed that some days were systematically pointed out as suspicious, even though they agree with the paper. We performed a subjective analysis in a case per case basis, and we classified further the values as *consistent* and *dubious*. The former were associated normally to heavy rain events concentrated in space; the latter show very different values even in very near stations. They were temporally removed from the database in order to not affect the PCA calculations (see González *et al.*, 1991).

The process ends when all dubious values coincide with paper. In table III the evolution of the depuration process is shown. In the first three stages, we only look for gross errors, checking essentially the scores associated with the leading patterns, which explains the low number of days affected.

Another important point is the measurement of efficiency. The column headed with η in Tables III and IV shows a number which even being independent of the number of stations seems to be pessimistic; in practice it is more representative the one indicated by column G (measured as errors per revised day), because in most cases the error was so obvious that by merely checking one or two our of the 21 values were enough to locate the error.

Stage	A	B	C	E	F	G	η
1	9	21	6	51			34
2	354	326	154	448		186	87
3	222	267	336	475	395	174	83
4	70	83	206	286	219	126	60
5	72	60	8	132	105	106	51
6	41	2	29	111	18	65	31
7	9	1	12	115	13	19	9
8			1	113	2	1	0
9				109		0	0

Table III: Evolution of the depuration of real errors for the full days (i.e. without missing values). We analyzed 21 stations. Keys to table: A.- Wrong values; B.- The digital value do not exist on paper; C.- Dubious value; E.- Total number of days checked; F.- Days not considered before; G.- (A+B+C)/E*100 Total number of errors for each 100 days revised; η= (A+B+C)/(21*E)*1000 Total number of errors for 1000 values checked.

Regarding the days with missing values, we applied the Penalty of Principal Components method (described by López *et al.* 1994). In those days with zero rain readings we simply assign zero to the missing values. In other case, we penalized the 10 weakest scores using as weights the reciprocal of the variance. Table IV shows the work in different stages, being all percentages to the total number of values revised in each stage.

Stage	A	B	C	D	E	F	G	η
1	344	314	220	945	457		399	210
2	56	27	65	57	495	94	41	22
3	117	118	138	37	558	179	73	39

4	52	69	118	21	536	94	49	26
5	17	36	36	10	586	53	17	9
6	21	12	34	6	560	30	13	7
7	19	20	9	1	659	15	7	4
8					659		0	0

Table IV: Evolution of the process of real errors for days with missing values. We analyzed 19 stations. Keys to table: A.- Wrong values; B.- The digital value do not exist on paper; C.-Dubious value; D.- Data exist on paper but were not digitized; E.- Total number of days checked; F.- Days not considered before; G.- (A+B+C+D)/E*100 Total number of errors for each 100 days revised; η=(A+B+C+D)/(19*E)*1000 Total number of errors for 1000 values checked

5. Conclusions

We have described and presented a methodology for multivariate quality control based upon Principal Component Analysis (PCA). The results, considering the effort involved can be regarded as satisfactory. In a controlled experiment we succeeded in identify one error every two days checked, finding that way over 80% of the known errors.

The required computer time can be considered minimal. The heaviest part is the calculation of the covariance matrix and its associated eigenvectors, an operation which is performed a limited number of times.

Considering that for each event it is only required a linear transformation, it is possible to apply the method in real time even with hand held computers.

6. Acknowledgments

Carlos López designed both the methodology and the experiment. Elizabeth González was in charge of the real errors phase, and also conducted the bibliographic search. Both authors jointly analyzed the obtained results. Jorge Goyret implemented in part the algorithms, and performed all calculations related with the described experiment.

It should be mentioned that this work is an extension of the task accomplished under the contract "Development of an hydrological model for the Río Negro catchment area" funded by UTE. The permission for using the information and publish the results is gratefully acknowledged.

7. References

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

DNM, 1988. "Procedimientos para el control de calidad climatológico" Informe interno de la Dirección Nacional de Meteorología, Nov. 1988, 20 págs.

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 data" In Proceedings of A workshop on ERS-1 wind and wave calibration, Schliersee, FRG, 2-6 June, 1986 (ESA SP-262, Sept. 1986)

Gnanadesikan, R.; Kettenring, J.R., 1972. "Robust estimates, residuals and outlier detection with multiresponse data" Biometrics, V 28, 81-124.

González, E.; Morales, C., 1991. "Depuración de la base de datos pluviométricos de la cuenca del Río Tacuarembó". Internal report prepared for the Hydrology Department of the Instituto de Mecánica de los Fluidos e Ingeniería Ambiental. 11 pp. (in spanish)

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.

Jácome Sarmento, F.; Sávio, E.; Martins, P.R., 1990. "Cálculo dos coeficientes de Thiessen em microcomputador". In Memorias del XIV Congreso Latinoamericano de Hidráulica, Montevideo, Uruguay (6-10 Nov., 1990). V 2, 715-724. (in portuguese)

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.

López, C.; González, J. F.; Curbelo, R., 1994. "Análisis por componentes principales de datos pluviométricos. b) Aplicación a la eliminación de ausencias" Estadística, 46, 146, 147, pp. 55-83

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.

Pio, C.A.; Nunes, T.V.; Borrego, C.S.; Martins, J.G., 1989. "Assesment of air pollution sources in an industrial atmosphere using principal components and multilinear regression analysis" The Science of the Total Environment, V 8, 279-292.

Richman, M.B., 1986. "Review article: Rotation of principal components" Journal of Climatology, V 6, 293-335.

Sevruk, B., 1982. "Methods of correction for systematic error in point precipitation measurement for operational use" World Meteorological Organization WMO 589, Operational Hydrology Report 21, 89 pp.

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" Final report. Part 2, Chapter 4. 83 pp. (in spanish)

Silveira, L.; Genta, J.L.; Anido Labadie, C., 1992a. "HIDRO URFING - Modelo hidrológico para previsión de caudales en tiempo real- Parte I: Simulación de los procesos hidrológicos en el suelo". Internal report 1/92 prepared for the Hydrology Department of the Instituto de Mecánica de los Fluidos e Ingeniería Ambiental, Facultad de Ingeniería, CC 30, Montevideo, Uruguay.

Silveira, L.; Genta, J.L.; Anido Labadie, C., 1992b. "HIDRO URFING - Modelo hidrológico para previsión de caudales en tiempo real- Parte II: Transformación en cuenca, ruteo y criterios de calibración y verificación" Internal report 2/92 prepared for the Hydrology Department of the Instituto de Mecánica de los Fluidos, Facultad de Ingeniería, CC 30, Montevideo, Uruguay.

White, D., 1991. "Climate regionalization and rotation of principal components" International Journal of Climatology, V 11, 1-25

Figures

Appendix 2

López, C., González, E. and Goyret, J., 1994, "Principal Component Analysis of pluviometric data b) Application to the missing value problem" *English translation of the paper* "Análisis por componentes principales de datos pluviométricos. b) Aplicación a la eliminación de ausencias" *Estadística (Journal of the Inter-American Statistical Institute) 46, 146,-147, pp. 55-83*

Principal Component Analysis of pluviometric data
b) Application to the missing value problem

English translation of the paper "Análisis por componentes principales de datos pluviométricos. b) Aplicación a la eliminación de ausencias" *Estadística (Journal of the Inter-American Statistical Institute) 46, 146-147, pp. 55-83.*

Carlos López, Juan F. González and Rosario Curbelo[1]

Abstract

The missing value problem is well known in all studies related either to natural phenomena or other areas. The present work has been motivated by the need to fill in the gaps in a daily pluviometric data bank in order to use it with an hydrological model. The spatial mean over the sub catchment area is calculated with the Thiessen's method, which on principle do not require a complete fill in. However, the method is highly sensitive to outliers in the case of few available measurements. The outlier detection problem has been analyzed in a companion paper, and here we will concentrate in reporting results for the missing value problem using some different methodologies. Such methodologies should preserve the main characteristics of the population, as well as the present quality and accuracy levels.

Results for four methods tested using a 15 year, daily pluviometric measurements are presented. The methods were the standard nearest neighbor, linear interpolation of the station time series, linear interpolation of the time series of the Principal Component Scores and Penalty of the Principal Component Scores. The last one were developed for this problem and proved to show the best behavior.

1. Introduction

According with Haagenson, 1982, Johnson, 1982, etc. objective analysis of both hydrological and meteorological fields are common practice. They are designed to interpolate an observed quantity using only sparse data. For the spatial mean rain field there exist other methods, like the Thiessen one (see Jácome Sarmento *et al*., 1990 for example) which might produce the required result without the need to imputate all missing values. Both situations led to a comparatively low interest in the research community, which was related with the scarce literature found in the specialized journals reviewed.

In the author's opinion in the overwhelming majority of the practical applications, the missing value is simply ignored, under the implicit assumption that those missing records appear at random, hypothesis which is rarely tested.

On the other hand, the topic is of major interest in statistics and social sciences; working group reports are mentioned in specific books, like the one of Rubin, 1987.

Of course somewhat sophisticated imputation methods do exist. For example, the one used at the US National Bureau of Census is quoted by Rubin, 1987. The idea is to imputate the missing value using a randomly selected one taken from the events which have an identical response in all the other answers (the method is originally designed for surveys). If there is

[1] Centro de Cálculo - Facultad de Ingeniería - Montevideo - Uruguay

Appendix 3

López, C., 1997, "Application of ANN to the prediction of missing daily precipitation records, and comparison against linear methodologies" *Third International Conference on Engineering Applications of Neural Networks. Stockholm, 16-18 June, 337-340.*

Application of ANN to the prediction of missing daily precipitation records, and comparison against linear methodologies[1]

Carlos López-Vázquez
Facultad de Ingeniería, Centro de Cálculo
CC 30, Montevideo, Uruguay
e-mail: carlos@fing.edu.uy

Abstract

Depending upon the user, weather records can be used as they are, or they need to be imputated prior its use. Despite the fact that general methods for meteorological variables exist, they are difficult to apply for daily rain. A specially difficult feature is that the overwhelming majority of the records (>80%) are of zero rain, leading to a very non-gaussian distribution. Other characteristic is the low autocorrelation of the time series.

The test region was the Santa Lucia river catchment area of 13000 km^2, at 35°S near the Atlantic; its yearly accumulated precipitation values are around 1000 mm, without a clear dry or wet season. The selected subset has 20 years long and 10 stations; 30% of the events show missing values.

A Monte Carlo simulation was designed, randomly choosing both date and station for the missing values and afterwards different imputation procedures were successively applied. Some statistics which characterize the distribution of the absolute error, namely its expected value, variance and 75, 85 and 95 percentile have been derived in order to compare the results.

Both traditional linear meteorological interpolation procedures as well as a suite of Backpropagation Artificial Neural Networks(ANN) has been compared. The present results are not very good, and show that is possible to imputate with a mean error of 2 mm/day and a RMS of 7 mm/day using both linear and nonlinear procedures, while ANN seems to be more robust against outliers.

Introduction

The problem of interpolate a field using sparse data is typical in many areas. In meteorology the objective analysis method (Haagenson, 1982; Johnson, 1982) is commonly applied because of its simplicity. It provides indirectly a way for calculating missing values using available data. However, the significant amount of information required by this method usually restricts its use to Global Data Assimilation Centers (Gandin, 1988), because they require historical records for the calculations.

Ideally the availability of all records is preferred, but there are meteorological problems which do not require a full dataset. For example, to calculate the areal mean value of rain the Thiessen-Voronoi polygons method (Jácome *et al.*, 1990) can be applied, without requiring extensive imputation of missing values.

[1] This work was funded by the Uruguayan CONICYT, under contract 51/94
Presented at the International Conference on Engineering Applications of Neural Networks. Stockholm, 16-18 June, 1997, pp. 337-340

Both situations led to a low interest on the topic reflected on the scarce meteorological papers on it. In most practical cases, either the missing value is ignored (assuming implicitly that the missing mechanism is random) or some ad-hoc technique is applied (linear interpolation, nearest neighbor, etc.) without further test or documentation. In any case, the population is clearly affected in an arbitrary way, under some hypothesis of unknown validity. However, it should be noticed that the missing value problem is of great interest in the Statistics and Social Sciences in general (Rubin, 1987).

Considered methods: a) linear

Due to its simplicity, this methods are widely used. They can be grouped together since the estimated quantity is a linear combination of the available data. Its general expression is $y_j = \underline{w}.\underline{x} + b$ being y_j the unknown quantity, $\underline{x}$ a vector with the available data and both the weight vector $\underline{w}$ and the number b are depending on the method. Typically vector $\underline{x}$ holds the values of the same day, and both $\underline{w}$ and b are constants for the whole dataset. This definition covers the methods of Cressman, Optimum interpolation (Gandin, 1965), Ordinary least squares, as well as other more simple ones, as the nearest neighbor. For the sake of completeness a brief description of them will follow:

- **Cressman**

The requested number is obtained after a linear combination with weights which are the inverse of square distance. The method does not require historical information, but only the station coordinates.

$$y_j = \frac{\sum\limits_{i \in N} y_i / d_{ij}^2}{\sum\limits_{k \in N} 1/d_{kj}^2} \text{ being vector } w_i = \frac{1/d_{ij}^2}{\sum\limits_{k \in N} 1/d_{kj}^2}\text{, and } b = 0$$

- **Optimum interpolation (Gandin, 1965; Johnson, 1992)**

This method is routinely applied for the initialization of global weather codes. Instead of interpolate the desired field, it interpolates the anomaly or difference with a simple predictor, and the spatial correlation properties of the anomaly field are analyzed. Usually it is assumed both isotropic and homogeneous, and it should be modelled in the general case. However, if the point where the prediction is required is one of the measuring point, its covariance with the other stations is available, and it looks very similar to the Ordinary Least Squares. The covariance might be calculated separately for winter and summer, for example, or used all together as we did. This procedure allows using information from the day before.

We used different anomaly fields and transformations for the variable to be interpolated which are summarized in table 1. For example, the method coded as "gandin7" assigns values for the variable $x_i = \sqrt{rain}$, taking the anomaly respect to its historical mean. In this case, $\underline{w}$ is fixed (following Johnson, 1992); $b = \overline{x}_j$ (the overbar stands for average over time). The classic Optimum Interpolation procedure is coded as "gandin20". Because daily rain has a very irregular probability density function (pdf) we designed a transformation $x = f(rain)$ which makes pdf(x) nearly uniform, except for $rain = 0$.

The transformation based on the cumulated density function is invertible and assures that x belong to the interval $[0,1]$.

Our coded name	*Anomaly*	*Variable to interpolate*	*Using data from days*	
	respect to:		*t*	*t-dt*
gandin	*historical mean*	*rain*	*X*	-
gandintrans	*historical mean*	*f(rain)*	*X*	-
gandin6	*historical mean*	*rain*	*X*	*X*
gandin7	*historical mean*	$\sqrt{rain}$	*X*	-
Initial value for the field chosen as zero				
gandin_diario	*0*	*rain-daily mean*	*X*	*X*
gandin4	*0*	*rain*	*X*	*X*
gandin5	*0*	*rain*	*X*	-
Neglecting instrumental error				
gandin20	*historical mean*	*rain*	*X*	-
gandin3a	*historical mean*	*rain-daily mean*	*X*	-

Table 1 Brief information about the different methods based on climatological functions. f(rain) denotes the transformation which renders a nearly uniform probability density function(see text). t and t-dt denotes values from the day and the day before

- **Ordinary Least Squares**

This method is completely standard and its theory can be found elsewhere. The weights $\underline{w}$ are chosen in order to minimize the 2-norm of the vector $M^{(j)}\underline{w} - m^{(j)}$ (a scalar proportional to the RMS) being $M^{(j)}$ the matrix of the available data (as many rows as dates, as many columns as stations but without the j-th one) and $m^{(j)}$ is a column vector with the j-th stations values. The implemented version assumes that the data is error free, so $\underline{w}$ can be derived from (dropping the index j) $M^T M\underline{w} = M^T.m$. The T stands for transpose. The number b is 0.

- **Least average (Least 1-norm)**

Here the weights $\underline{w}$ are chosen in order to minimize the 1-norm (sum of absolute values) of the vector $M^{(j)}\underline{w} - m^{(j)}$. This is a much more difficult problem because it does not lead to a linear system of equations and has to be solved as a non-linear optimization task. Also it requires substantially more CPU time than all previous methods.

- **Least 95 percentile**

Since the population might be affected by a small set of gross errors, it is fit to minimize a robust statistic, as the 95 percentile of the distribution of errors. As before, this problem requires significant CPU time.

- **Nearest Neighbor**

We considered two criteria for the distance: euclidean and qualitative similarity. In both cases the missing value is taken directly from another station following a given order. In the first case, the order is due to geometrical distance, and in the second we used the expertise from a meteorologist. All weights are zero, except one which is 1, and the number b is 0.

- **Assign a constant value**

This is a simple method, which disregard any other information. We applied it using the modal value and the expected value. For our dataset, the former is 0 mm/day and the latter is near to 3 mm/day.

Considered methods: b) Non linear methods (ANN)

Such methods are very new and they are based upon simple models of the biological neural networks. They have been used for the short term prediction of SO_2 concentration (Boznar *et al.,* 1993), electrical load (Park*.,* 1991), etc. The ANN is organized in layers, being the first one stimulated directly by the observed values; each neuron of the next layer is stimulated by a linear combination of the outputs of the previous layers by means of a simple transfer function, like the logsig (Demuth *et al.,* 1994) given by:

$out_j = \left\{ 1 + \exp\left[-\sum_i \left(a_{ij} * input_i \right) \right] \right\}^{-1}$, with parameters a_{ij} to be fixed for each neuron. The ANN requires, as its biological counterpart, a training process which is simulated here by means of adjusting the a_i parameters. In this work we compared one two-layer net, with 6 logsig neurons in the hidden and 1 linear neuron in the output layer, and one three-layer one, with 8 linear neurons, 4 logsig and one logsig for the output. Both were trained using one third of the available values trying to minimize the RMS of the error. The error is defined as the difference between ANN output and true value. For the first case we substracted rain values in mm/day while for the second case something different has to be done, since the last neuron has an output belonging to the interval $[0,1]$. We trained the net in order to minimize the error with the transformed rain $x = f(rain)$. All nets were trained using *backpropagation* (Rumelhart *et al*., 1986) and due to practical reasons the number of iterations was kept low, so its performance might be improved with more iterations. Its training cost in CPU time is high: over 10 hours of SUN 20 for each meteorological station.

Conclusions and results

After 250 simulations, the results are summarized in table 2. It should be noticed the improved results for those methods using information from the day before (gandin4, gandin6 and gandin_diario). Among those which use only information of a single day, the best results are obtaining by the minimum 95 percentile, closely followed by the Ordinary Least Squares method.

It should be stressed that, since the database still has errors, it is possible that the methods suggest suitable values and the outliers affect some of the considered statistics. This is unlikely to occur for the 85, 95, etc. percentile, and then the importance of the ANN denoted bp7.

As final conclusions:

a) common methods based upon mere substitution by a neighbor or by a constant gave poor results.

b) as expected, optimum interpolation methods outperforms the others in terms of RMS, fairly close to the ordinary least squares and least 95 percentile.
c) non linear methods (very expensive in the training phase) led to slightly more robust results, but renders similar figures in terms of average and RMS.

	Average	75%	85%	95%	RMS
bp1	2.65	1.92	4.53	**13.03**	**7.15**
bp7	2.51	1.28	3.64	**12.54**	7.71
cressman	2.63	**0.80**	4.58	15.75	8.20
gandin	2.64	1.48	4.20	13.57	7.24
gandin3a	2.60	1.83	4.73	13.96	7.42
gandin20	2.68	1.56	4.21	13.42	7.21
gandin4	2.53	1.92	4.59	13.28	**7.02**
gandin5	2.39	1.25	4.14	13.64	7.25
gandin6	2.71	2.06	4.72	13.39	**7.05**
gandin7	**2.23**	**0.50**	**3.11**	13.39	7.48
gandin_diario	**2.04**	0.89	**2.99**	**11.01**	7.66
gandintrans	3.06	**0.80**	4.52	13.46	8.11
least squares	**2.34**	1.33	**4.09**	13.13	**7.01**
least 95's percentile	**2.34**	1.34	**4.10**	**13.07**	**7.01**
least average	**2.26**	0.86	3.60	**13.23**	7.21
modal value	2.79	**0.00**	**1.78**	19.04	10.26
expected value	4.73	2.96	**3.02**	16.25	9.88
geometrical distance	2.76	**0.02**	4.22	17.37	9.13
expert distance	2.82	**0.01**	4.31	17.74	9.33

Table 2 Preliminary results in mm/day for the different imputation methods. The expected value and the 75, 85 and 95 percentile of the distribution of the absolute error, and its RMS are presented and compared. In bold the five best results for each estimator.

References

Boznar, M.; Lesjak, M. and Mlakar, P., 1993 "A neural Network-based method for short-term predictions of ambient SO2 concentrations in highly polluted industrial areas of complex terrain" Atmos. Environ., V 27B, N 2, 221-230

Demuth, H. and Beale, M. 1994 "Neural Network User's guide (Toolbox for MATLAB)" The MathWorks, Inc. 226 pp., http://www.mathworks.com

Gandin, L. M., 1965 "Objective analysis of Meteorological Fields". Israel Program for Scientific Translations, 242 pp.

Gandin, L. M., 1988 "Complex Quality Control of Meteorological Observations". Mon. Wea. Rev, V 116, 1137-1156

Haagenson, P.L, 1982 "Review and evaluation of methods for objective analysis of meteorological variables" Papers in Meteorological Research, V 5, N 2, 113-133.

Jácome Sarmento, F.; Sávio, E. and Martins, P.R., 1990 "Cálculo dos coeficientes de Thiessen em microcomputador". In Memorias del XIV Congreso Latinoamericano de Hidráulica, Montevideo, Uruguay (6-10 Nov., 1990). V 2, 715-724 (in portuguese)

Johnson, G. T. 1982 "Climatological Interpolation Functions for Mesoscale Wind Fields". Journal of Applied Meteorology, V 21, N 8, 1130-1136

Park, D. C., 1991 "Electric load forecasting using an artificial neural network" IEEE Transactions on Power Systems, N 2, 442-449

Rubin, D. B., 1987 "Multiple imputation for nonresponse in surveys". John Wiley and Sons, 253 pp.

Rumelhart, D. E.; Hinton, G. E. and Williams, R. J. 1986 "Learning representations by Back-Propagating errors", Nature, V 323, 533-536

no other event in such condition, a distance between surveys is defined, and the nearest is chosen for imputation.

Another typical and simple method is to make a regression over the dataset, fitting a mathematical model. Usually partial or total least squares as well as principal components are used, as presented by Stone *et al.*, 1990.

All the abovementioned methods produce a single value for a each missing value. Quoting Rubin, 1987 "... in general is intuitive that imputating using the *optimal* value for each missing value will underestimate variability...". However, the possibility to obtain more than a single value for each missing one can be considered. Rubin, 1987 described a set of techniques (some too much specialized to surveys). The general idea is create, for each missing value, m possible alternative values (with m small) and considering that m different complete sets are available. If the missing value rate is low, the method might be of use, requiring however more space (for saving the multiple imputations) and also more computation time (for processing separately the different sets). For further details please see Rubin, 1987.

2. The present work

2.1 Motivation

This work can be considered as a natural extension of the preliminary treatment of the pluviometric data used in the calibration phase of a flow-rain-flow hydrological model for the Río Negro catchment area. Three hydropower dams operate sequentially there, operated by the national electrical utility (UTE[2]). For further details please refer to Silveira *et al.* (1991, 1992a y 1992b).

2.2 General characteristics of the study area

a) Geographical

Even though we have analyzed a greater catchment area, we restrict ourselves for this paper to the Río Tacuarembó catchment area, of about 20.000 km^2, located at 32° S 55° W, at 400 km from the ocean. The typical landscape is smooth, with heights below 500 m, few canyons and lakes. The typical monthly rain rate is within 74 and 120 mm/month.

b) Measuring net

The national net defined by the DNM[3] is based on a regular grid of size 10 km, sequentially numbered. The identification for the station is the same of the cell which it belongs; from time to time two or more stations might be operating within the same cell, so a letter A, B etc. is appended to the identification code. We coined the term *synonym* for those stations. For our purposes, we will consider all the stations belonging to the same cell as the same station. From the administrative point of view, our net is the sum of four ones, independently operated by different institutions. Those nets have different spatial density and reliability. Its structure has been changing during time, and any station might:

- Start operating in any moment during the period

[2] UTE - Administración de Usinas y Transmisiones Eléctricas

[3] DNM - Dirección Nacional de Meteorología

- Stop operating in any moment during the period
- Start operating to substitute another one which has been withdrawn
- Be replaced by other or not.

For our purposes we will only distinguish those stations that use to be operated by AFE[4], which systematically adds the readings from Sunday and Monday and consider them as belonging to Monday. In the abovementioned catchment area 21 stations are in operation, and we selected 13 for this work.

c) Dataset

As mentioned before, the topology of the net usually suffer from transformations. According to Silveira et al., 1991 it exists at present too many stations. Since many of them have synonyms we join their records and disregard any distinction within a cell.

For this paper we selected a subset of 13 stations, located as shown in fig. 1, which have been carefully checked for typing errors by using some algorithms presented in López *et al.*, 1994. We restrict ourselves to records from Jan 1st. 1975 to Dec 2nd. 1989, covering nearly 15 years.

3. Methods used for the Missing value problem

3.1 Nearest neighbor

The method assigns a list of alternative stations to the one which is intended to imputate; the missing value will be replaced by a number taken from the first one with measurements for the particular date. On principle the abovementioned list is ordered according to increasing distance to the one intended to imputate; however some confidence considerations are taken into account and the purely geometric order might be altered.

3.2 Time series interpolation

If the value for the day t_f and station j is missing, we search for the nearest previous and following reading available at station j, and a simple linear interpolation is performed.

Let us denote as t_f the date of the missing record and as $p_j(t_f)$ the unknown value for the day t_f and station j.

Let us denote also as t_{f-m} the latest day before t_f with readings, and as t_{f+r} the first day after t_f with readings $\left(t_{f-m} < t_f < t_{f+r}\right)$. The interpolation rule for the missing value is

$$p_j(t_f) = p_j(t_{f-m}) + \frac{t_f - t_{f-m}}{t_{f+r} - t_{f-m}}\left(p_j(t_{f+r}) - p_j(t_{f-m})\right) \qquad (1)$$

[4] AFE - Administración de los Ferrocarriles del Estado

3.3 Time interpolation of the Principal Component Scores series (TIPS)

This method is based upon Principal Component Analysis (PCA), which have been presented in the companion paper by López *et al.*, 1994. We will present briefly the notation and refer the reader to the abovementioned reference.

Let us name as $\mathbf{P}_{(n,1)}(t)$ the precipitation vector of the n selected stations for the time t. Let's define a rectangular matrix $\mathbf{M}$ which rows are the vectors $\mathbf{P}(t_{m_j}) - \mathbf{P}_M, j = 1..r$, defined for those days without missing values. $\mathbf{P}_M$ is the mean vector for the considered period.

The eigenvectors of matrix $\mathbf{C}_{(n,n)} = \mathbf{M}^T * \mathbf{M}$ are named principal components or patterns, and will be denoted as $\mathbf{e}_i$. We will assume that the associated eigenvalues are ordered, and decrease with i. The relationship between the pluviometric records $\mathbf{P}_{(n,1)}(t)$ and the scores represented as the vector $\mathbf{A}_{(n,1)}(t)$ is

$$\mathbf{P}(t) = \mathbf{P}_M + \mathbf{E.A}(t) \qquad (2)$$

where $\mathbf{P}_M$ stands for the mean vector for the period and $\mathbf{E}_{(n,n)}$ is the matrix formed by the eigenvectors $\mathbf{e}_i$.

$$\mathbf{P}(t) = \begin{bmatrix} \mathrm{p}_1(t) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \mathrm{p}_\mathrm{n}(t) \end{bmatrix} ; \mathbf{P}_M = \begin{bmatrix} \overline{\mathrm{p}_1} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \overline{\mathrm{p}_\mathrm{n}} \end{bmatrix} ; \mathbf{A}(t) = \begin{bmatrix} \mathrm{a}_1(t) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \mathrm{a}_\mathrm{n}(t) \end{bmatrix} . ; \mathbf{E} = \begin{bmatrix} \\ \mathbf{e}_1\, \mathbf{e}_2 \,..\, \mathbf{e}_n \\ \\ \end{bmatrix}$$

If $\mathbf{C}_{(n,n)}$ is non-singular then $\mathbf{E}_{(n,n)}$ is invertible, so given the readings $\mathbf{P}(t_{f-m})$ y $\mathbf{P}(t_{f+r})$ it is possible to obtain vectors $\mathbf{A}(t_{f-m})$ and $\mathbf{A}(t_{f+r})$ from (2).

Eq. (2) can be expressed as well as

$$\mathbf{P}(t) = \mathbf{P}_M + \sum_{\mathrm{i=1}}^{\mathrm{i=n}} \mathbf{a}_\mathrm{i}(t).\mathbf{e}_i \qquad (3)$$

For any intermediate time $t_l, l \in \left(f - m + l, f + r - 1\right)$ the precipitation is calculated by linear interpolation of vector $\mathbf{A}(t)$. On principle, all the readings for time t can be obtained from eq. (2).
By analyzing the scores a_i it is clear that the standard deviation of a_i decreases as i increases making typically a minor contribution to the summation.
It is then a natural conclusion that for the reconstruction of vector $\mathbf{P}(t)$ all terms with i>q can be neglected for some q, without substantial loss of information. Then the equation (3) can be substituted by the following approximate expression

$$\mathbf{P}(t) = \mathbf{P}_M + \sum_{\mathrm{i=1}}^{\mathrm{i=q}} \mathbf{a}_{\mathrm{i}}(t).\mathbf{e}_i \tag{4}$$

Summing up, if there is at least one missing value of vector $\mathbf{P}(t_f)$ corresponding to time t_f we search for the nearest previous and following without missing values. It should be stressed that, in opposition with the standard linear interpolation, this method uses all n stations, and not each one independently.
Let denote as t_f the day to be imputated. Let also be t_{f-m} the nearest previous day without missing values and t_{f+r} the nearest following $\left(t_{f-m} < t_f < t_{f+r}\right)$. Since both t_{f-m} and t_{f+r} are complete days the scores $\mathbf{A}(t_{f-m})$ and $\mathbf{A}(t_{f+r})$ corresponding with vectors $\mathbf{P}(t_{f-m})$ y $\mathbf{P}(t_{f+r})$ can be easily calculated using eq. (2).
Then, for time t_{f-m+l}, the vector of scores $\mathbf{A}(t_{f-m+l})$ can be calculated by linear interpolation of both vectors $\mathbf{A}$ mentioned before. The first guess for the precipitation of time t_{f-m+l} can be obtained from eq. (4).
However at time t_{f-m+l} there are some values recorded. The missing ones can be taken from the first guess vector, and then we can complete all elements vector $\mathbf{P}(t_f)$ using as much available information as possible.
Once completed the time t_f we are in position for a new interpolation, using now vectors $\mathbf{P}(t_{f-m+l})$ and $\mathbf{P}(t_{f+r})$ as starting point; this step can be repeated as many times as necessary, in order to fill all the gaps.
The performance of this approximation is heavily connected with the autocorrelation properties of the scores a_i. For meteorological variables this autocorrelation properties are very different for different a_i, which is another argument to limit the number of terms in eq. (4).
For example, in the work of Cisa *et al.* (1990) it is shown that for the hourly surface wind in southern Uruguay the time lag T_i required for the autocorrelation of the i-th score time series to take for the first time the value 0.5 is (25,9,5,3,3,...,1.2,1) for i=1...15, being the bigger values for the most important PC. T_i is measured in hours.

Such situation is not the case for daily rain, because all scores have a dramatic drop for T_i=l day, being 1 day the sampling period (see figs. 2 and 3). This fact explains in part the poor results obtained with this method, despite it is better than the one obtained with the standard linear interpolation of station time series.

3.4 Penalty of the Principal Component Scores

If we analyze the histogram of the scores a_i it can be observed that for the main PC it is heavy skewed (asymmetric?), or has a significant dispersion around zero. On the other hand, for the weak PC the histogram is symmetric and the dispersion around zero is very low. As an example see figs. 4 and 5. They have been obtained from a slightly modified population, because all events with null precipitation *in all the stations* have been removed for this plot. Any imputation procedure should preserve this properties, and then it should produce scores a_i consistent with this histograms, i.e. very near zero for all weak PC. Such property might be imposed as a condition, choosing for any given date all missing components of vector $\mathbf{P}(t_f)$ in order to minimize some penalty function, like

$$S(\mathbf{P}) = \sum_{i=k}^{i=n} w_i . a_i^2(\mathbf{P}) \qquad (5)$$

being the scores $a_i(\mathbf{P})$ corresponding to the vector $\mathbf{P}$ (now complete) and the weights w_i selected in order to consider the different absolute value of each score a_i. Vector $\mathbf{P}$ is only partially known, and it is assumed that it has q unknowns (or missing values). The optimum of S can be obtained making its partial derivatives null for all unknowns

$$\frac{\partial S}{\partial p_{m(j)}} = 0, \;\; \mathrm{j} = 1..\mathrm{q}$$

being $p_{m(j)}$ the missing records for this time. The so defined linear system can be easily solved by standard procedures.

4. Experimental procedure

We generate at random pairs of time-station which will be regarded as fictitious missing values. Valid pairs are those which have been measured; all methods should calculate a value to *imputate* it. We will denote as *real value* the measured one, and *calculated value* as the one obtained by means of an imputation procedure. If the pair is not valid, it is simply discarded.

Once a prescribed number of valid pairs have been processed the standard deviation of the difference between real and calculated values for each method is calculated; it will be the main statistics to compare within methods.

The total number of pairs is taken as a percentage of the days in the analyzed period (5450 days from Jan 1st. 1975 to Dec. 2nd. 1989) and we restrict the number of missing values per day to one. We varied the percentage from 20% through 80% and no significant difference in the results were noticed.

For all methods we made runs considering all pairs; for the nearest neighbor and Penalty we made also runs disregarding those events with zero rain in all stations, which account for 80% of the cases. We did it as an attempt to avoid the negative impact of such significant amount of constants in the estimators. The results in terms of standard deviation increased two times approximately, but the relative values for different methods remained the same.

We depurated as much as possible the data bank against the original paper records. All 13 stations show less than 5% missing values for the period, which we considered enough for the purposes of this work.

We also made some sensitivity checks for the parameters of the different methods: for both methods using scores, we varied the number of terms to consider; for the nearest neighbor method we varied the list of alternatives (by removing the nearest ones).

5. Results

All figures correspond with an experiment with 2091 days with missing values, implying 53% of the analyzed days approximately.

5.1 Nearest neighbor method

Some calculations using only the list of 13 stations were performed; also we enlarged the set by using up to 86 stations located in the catchment area as well as in the vicinity. The precedence order in this case was strictly geometric distance.

We confirmed that the availability of a dense network of stations in a regular topography like this improves the results. If we delete alternatives from the list increasing the mean distance between the station to be imputated and its alternatives is clear (see fig. 6) that the standard deviation increases.

The regularity characteristics of the rain phenomena in this region even up to distances of 150 km apart leads to good results. This mean distance is defined here as the expected value of the geometric distance weighted by the frequency of substitutions by an alternative. The maximum distance between any of the 13 stations is 201 km.

The point of fig. 6 with mean distance 33.7 km results from choosing alternatives from the set of 13 stations. The resulting standard deviation is $\sigma = 5.55$ mm/day which has to be taken into account when comparing the different procedures.

In general the alternative stations have not been corrected at all; one consequence is that even with a lower mean distance, the alternative set renders a bigger standard deviation than the abovementioned case of the 13 stations.

5.2 Linear interpolation of the station time series

As mentioned before, good results can be expected for this method if the considered phenomenon evolves slowly in relation with the sampling interval, or if any contiguous gap is shorter than the typical time scale of the problem.

The rain in Uruguay shows significant variation in time, and generally the meteorological event occurs under the form of storms more or less concentrated in time (from hours up to three or four days). That's why using daily sampling this method is unlikely to produce good results.

The results for the set of 13 stations show a standard deviation of 12 mm/day.

5.3 Linear interpolation of the time series of the Principal Component Scores

As expected this method renders similar results as the standard interpolation of the time series. Despite it requires more CPU time, the results are not significantly better: the standard deviation of the error varies from 11.3 and 11.83 mm/day depending on the number of terms using in the calculation.

The weak effect in using more or less terms in the calculations is not surprising, and it is related with the extremely low time autocorrelation of the phenomena. In fig. 7 the evolution of the standard deviation respect to the index q (see eq. 4) is shown.

It is expected that this method will produce significantly better results for other meteorological variables - like surface temperature, surface wind, etc. - but this extreme have not been tested yet.

5.4 Penalty of the Principal Component Scores

This method renders the best values in terms of the standard deviation, with a minimum value of 4 mm/day. To give an indication of the variability of the time series itself we calculated its standard deviation. For any station its value is over 15 mm/day for the whole period.

The value of 4 mm/day has been obtained by using k in the range 8 to 10 (see eq. 5) which implies 3 to 5 terms in the equation. For example, if k is equal 1 all the scores are taken into account, so the surface is forced to resemble the historical mean value. For k near n, only the weakest PC are affected in the summation S, but some others which also explain noise are left uncontrolled. This fact explains why the results are of poor quality.

The task of fixing the optimum k for each data bank can be accomplished by means of an experiment like this, or by a subjective analysis of the PC. The shape of the isolines might easily distinguish from those which are related with noise from those which are related with the physics. The weakest PC are also very sensitive to outliers (see Silveira *et al.* 1991).

From fig. 8 it can be analyzed the dependence of the standard deviation of the error in relation to the number of terms analyzed. It is clear that the optimum is robust in relation with k.

The weights w_i were chosen in order that all terms $w_i a_i^2$ are of comparable order. In early stages we adopted as w_i the inverse of the RMS. of the time series of the scores a_i. Despite being reasonable this rule revealed soon unsuitable for the main patterns because they are clearly asymmetric. Thus, we defined the weights as $w_i = 1/\alpha_i^2$, being the limit α_i defined in order to

$$\int_{-\alpha_i}^{\alpha_i} a_i^2 . f_i(a).da > 0.96$$

being f_i the probability distribution function for the score a_i , i=1..n. It is automatically verified that for less than 4% of the events, $w_i . a_i^2(t) \geq 1.0$

6. Conclusions and recommendations

From the results presented it is clear that all methods which rely on the temporal behavior of the rain might distort significantly the general characteristics of the population, in particular if the proportion of missing value is important.

Both the Nearest neighbor and Penalty of principal component scores methods which take into account the spatial behavior of the rain render significantly better results than the ones based on temporal properties. The main reasons are the characteristics of the rain in terms of the time sampling strategy and the smooth topography of a small catchment area.

The method named Penalty of Principal Component scores shows a standard deviation 28% less than the Nearest neighbor one (4.01 vs. 5.55 mm/day), a value which has been obtained using only measurements of 13 stations.

If one includes more stations, the standard deviation is even bigger, even for lower mean distances, which can be partly explained by the unknown quality of the extra stations.

Another advantage for this proposed method is the modest computer resources involved. For instance, even a hand held computer might be enough, since the only requirements is that it can hold a matrix of size n, a vector of size n for the mean averages, and the capabilities to solve a linear system of equations. This feature should be considered for routine operation of any hydrologic model.

As a further improvement of the procedure, we plan to minimize the joint probability of the scores a_i which in turn implies the solution of a non lineal problem for each event with missing values. Despite a (significantly) increased computer time cost this procedure is theoretically more sound.

7. Acknowledgments

Juan González implemented and performed the calculations for both the nearest neighbor and temporal interpolation of the station series, and Rosario Curbelo programmed the linear interpolation of the time series of the Principal Component Scores. Carlos López developed and coded the Penalty of the Principal Component Scores program. The methodological and experimental design was due to Carlos López. All three authors took part in the discussion of the results. It should be mentioned that this work is an extension of the task accomplished under the contract "Development of an hydrological model for the Río Negro catchment area" funded by UTE. The permission for using the information and publish the results is gratefully acknowledge.

8. References

González, E.; Morales, C., 1991. "Depuración de la base de datos pluviométricos de la cuenca del Río Tacuarembó". Internal report prepared for the Hydrology Department of the Instituto de Mecánica de los Fluidos e Ingeniería Ambiental. 11 pp. (in spanish)

Haagenson, P.L, 1982. "Review and evaluation of methods for objective analysis of meteorological variables" Papers in Meteorological Research, V 5, N 2, 113-133.

Jácome Sarmento, F.; Sávio, E.; Martins, P.R., 1990. "Cálculo dos coeficientes de Thiessen em microcomputador". In Memorias del XIV Congreso Latinoamericano de Hidráulica, Montevideo, Uruguay (6-10 Nov., 1990). V 2, 715-724.

Johnson, G.T. 1982. "Climatological Interpolation Functions for Mesoscale Wind Fields". Journal of Applied Meteorology, V 21, N 8, 1130-1136.

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.

López, C.; González, E.; Goyret, J., 1994. "Análisis por componentes principales de datos pluviométricos. a) Aplicación a la detección de datos anómalos" Estadística, V 6, N 146-147, 55-83.

Richman, M.B., 1986. "Review article: Rotation of principal components" Journal of Climatology, V 6, 293-335.

Rubin, D. B., 1987. "Multiple imputation for nonresponse in surveys". John Wiley and Sons, 253 pp.

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" Final report. Part 2, Chapter 4. 83 pp. (in spanish)

Silveira, L.; Genta, J.L.; Anido Labadie, C., 1992a. "HIDRO URFING - Modelo hidrológico para previsión de caudales en tiempo real- Parte I: Simulación de los procesos hidrológicos en el suelo". Internal report 1/92 prepared for the Hydrology Department of the Instituto de Mecánica de los Fluidos e Ingeniería Ambiental, Facultad de Ingeniería, CC 30, Montevideo, Uruguay.

Silveira, L.; Genta, J.L.; Anido Labadie, C., 1992b. "HIDRO URFING - Modelo hidrológico para previsión de caudales en tiempo real- Parte II: Transformación en cuenca, ruteo y criterios de calibración y verificación" Internal report 2/92 prepared for the Hydrology Department of the Instituto de Mecánica de los Fluidos, Facultad de Ingeniería, CC 30, Montevideo, Uruguay.

Figures

Appendix 4

López, C., 1997, "Locating some types of random errors in Digital Terrain Models" *International Journal of Geographic Information Science, **11**, 7, 677-698*

Locating some types of random errors in Digital Terrain Models[1]

Carlos López
Environmental and Natural Resources Information Systems
Royal Institute of Technology
Stockholm 100 44, Sweden[2]

Abstract:
The increasing use of Geographic Information System applications has generated a strong interest in the assessment of data quality. As an example of quantitative raster data, we analyzed errors in Digital Terrain Models (DTM). Errors might be classified as systematic (strongly dependent on the production methodology) and random. The present work attempts to locate some types of randomly distributed, weakly spatially correlated errors by applying a new methodology based on Principal Components Analysis. The Principal Components approach presented is very different from the typical scheme used in image processing. A prototype implementation has been conducted using MATLAB, and the overall procedure has been numerically tested using a Monte Carlo approach. A DTM of Stockholm, with integer-valued heights varying from 0 to 59 m has been used as a testbed. The model was contaminated by adding randomly located errors, distributed uniformly within -4m. and +4m. The procedure has been applied using both spike shaped (isolated errors) and pyramid-like errors. The preliminary results show that for the former, roughly half of the errors have been located with a type I error probability of 4.6% on average checking 1 per cent of the dataset. The associated type II error of the larger errors (of exactly +4m. or -4m.) drops from an initial value of 1.21% down to 0.63%. By checking another 1 per cent of the dataset such error drops to 0.34% implying that about 71% of the ±4m errors have been located; type I error was below 11.27%. The results for pyramid-like errors are slightly worse, with a type I error of 25.80% on average for the first 1 per cent effort, and a type II error drop from an initial value of 0.81% down to 0.65%.

The procedure can be applied both for error detection during the DTM generation and by end users, and it might be of use for other quantitative raster data examples.

I Introduction:

Data quality has become an important aspect of Geographic Information Systems (GIS) applications. John (1993) stated that "...very wrong answers can be derived using perfectly logical GIS analysis techniques, if the user is not aware of the particular peculiarities of their data..."

Although this statement holds for any kind of data, we will concentrate here on the case of Digital Terrain Models (DTM). We will not consider errors in the intermediate steps in the process of DTM generation, but we will concentrate on the errors in the final product..

Östman (1987) pointed out the fact that there exists no unique criteria or single measure for the "quality" of a DTM. He suggested that at least, one should consider accuracy in height,

[1] Published in International Journal of Geographical Information Science, 11, **7**, 677-698, 1997

[2] Permanent address: Centro de Cálculo, Facultad de Ingeniería, Universidad de la República, CC 30, Montevideo, Uruguay

slope and also curvature. In his paper, the performance of an "on line" editor is described. It attempts to find gross-errors while the DTM is being created. This editor was intended to correct mainly those errors that affect curvature or slope, so no substantial ability to improve the height accuracy is reported. He pointed out that gross errors typically account for less than 0.5% of the whole dataset.

Day *et al.* (1988), tested three methods for the generation of DTM based on SPOT data. The three results were compared with a very carefully, manually digitized 30 m. grid DTM, in terms of height differences. Even though the goal of the work was to compare the operational behavior of the algorithms, their paper does not propose any solution for the locations of the errors. The distribution function of the absolute size of such errors is also presented for each method. Similar results arises from the work of Theodossiou *et al.* (1990).

Most of the literature concentrates on the location of gross errors in early stages of the production process. Bethel *et al.* (1984) proposed the method of maximum chi-squared ratio for on line quality control. A dense regular grid of observation is assumed. The method is restricted to the detection of gross errors in the image matching process, and is based upon the hypothesis that a least-square adjustment with bicubed spline function may fit locally a DTM. The occurrence of big residuals suggest the existence of a blunder. He tested the methodology using spike-like blunders of no more than 10 feet (about 3 m).

Even though the goal of this work is not to analyze the different ways a DTM can be produced, Ackermann, (1995) points out that the trend in DTM production is towards a move from interpolation to aproximation, because the new generation equipment is able to produce many height values, but possibly with less accuracy than traditional equipment. The surface is aproximated using many points, instead of being interpolated from few carefully obtained values.

In a review of general statistical methods, Barnett *et al.* (1984) classify the current methodologies for error detection in two classes, provided the distribution of the variable is known a priori, or not at all. The first class includes methods that typically require also the estimation of the parameters of the distribution. They are unlikely to work properly here.

For DTM applications, only those methods which do not require any particular distribution (also called Informal Methods) are suitable. Such methods include techniques related to Cluster Analysis (see for example, Fernau *et al.*, 1990), graphical methods, Principal Component Analysis, and others. Some of them were originally developed for applications in social sciences, but are increasingly used in other fields.

Principal Component Analysis (PCA) is a widely known technique, both in digital image processing and in the treatment of time series. Its ability to extract uncorrelated patterns that enhance the interpretability of the data, and the possibility to reduce the number of patterns separating the physics from the noise is well known.

In the field of remote sensing, principal components analysis is used to reduce the number of image bands of information (Chavez et al., 1989; Eklundh et al., 1993). Essentially, the remote sensing image data is re-mapped into a new coordinate system reducing the dimensionality of the data. For example, rather than analyzing data from 7 Thematic Mapper bands, we can do a principal components analysis to reduce the number of image bands to 3 or 4 bands of information that contain most of the variance of the data. Normally, we disregard the information in principal components beyond the third band, as this information relates to noise in the data set. This direct approach is not suitable for DTM analysis since at most a single model is usually available.

Summing up, most of the literature uses tailored procedures to locate gross errors. They concentrate mostly on the DTM production stage, disregarding the problem faced by the end user. This paper presents a new methodology for locating not merely gross errors but also some subtle ones, which can be applied both by the producer and the end user. The methodology were tested in a real DTM using numerically simulated errors, and the results are presented.

The paper is organized in eight sections. In section II a description and quick introduction to the PCA technique is sketched. Section III introduces the methodology for elongated DTM. In section IV the proposed procedure is described in terms of as a step-by-step recipe. Section V describes the Monte Carlo experiments designed to test the methodology. Section VI show the results in a particular DTM test case, using different error shapes. Finally, section VII contains a discussion and sectionVIII is devoted to conclusions where the results and proposed future work are discussed. Acknowledgements and References are included under headings IX and X.

II Principal Component Analysis in brief

The theory of PCA can be found in many textbooks, for example, Lebart *et al.* (1977). To make clear both the notation and the terminology, a brief sketch of the major concepts and results will be presented.

Given a table of n events of w variables, they can be represented as n points in the R^W space. They are supposed to be homogeneous , i.e., share the same measuring units. Each k-point (event) corresponds to a point in R^W, and each event is composed of w scalar observations. The case of w=3 is illustrated in the figure 1, where each point M_k represents an event. The PCA attempts to find the direction $\mathbf{e}_1$ of the vector in R^W space which minimizes the sum of distances M_kH_k squared, taken over all k (see fig. 1). The origin O is the centroid of the set of points. For the sake of clarity in the figure, points with negative coordinates are not shown.

The projection OH_k, which is also the scalar product of vector $\mathbf{M}_k$-$\mathbf{O}$ with the unitary vector $\mathbf{e}_1$, is called here the score (following Richman, 1986). Thus $\mathbf{M}_k$-$\mathbf{H}_k$ is orthogonal to $\mathbf{e}_1$. There is one score value associated with vector $\mathbf{e}_1$ for each point in R^W. Let us also assume that $\mathbf{e}_1$ is unique.

If all the values M_kH_k are zero, we have reduced the problem of original dimension w, to a one-dimensional one. All the variability in the observations is explained by a single vector $\mathbf{e}_1$. If this is not the case, we may try to repeat the procedure with the remaining variability M_kH_k, which belongs to a (w-1) subspace of R^W orthogonal to $\mathbf{e}_{1.}$ The original measurements $\mathbf{M}_k$ - $\mathbf{O}$ can be replaced with the difference $\mathbf{OM}_k$ - $\mathbf{OH}_k$, which is equal to $\mathbf{M}_k$ - $\mathbf{H}_k$.

There should be a vector $\mathbf{e}_2$ (automatically orthogonal to $\mathbf{e}_1$) which minimizes the distance in the R^W space. The process continues the same way, being each new vector $\mathbf{e}_p$ orthogonal to all the previous ones, and there are w such vectors. Those vectors are called principal components (PC).

Each event $\mathbf{M}_k$ - $\mathbf{O}$ can be expressed as a linear combination of the PC

$$\mathbf{M}_k - \mathbf{O} = a_1(k) * \mathbf{e}_1 + a_2(k) * \mathbf{e}_2 + a_3(k) * \mathbf{e}_3 + ... + a_w(k) * \mathbf{e}_w \qquad (1)$$

It can be shown that the scores $a_i(k)$ associated with vector $\mathbf{e}_i$ are uncorrelated with those of vector $\mathbf{e}_j$. The vectors $\mathbf{e}_i$ are the eigenvectors of the covariance matrix of the data, and its components are named *loadings* in the literature. The sum of the corresponding eigenvalues equals the sum of the squares of the distances M_kH_k (Lebart *et al.*, 1987).

PCA analysis renders a sequence of principal components, which explains most (or all, for p=w) of the variance of the data. That implies that the error in approximating the data with a linear combination of their first p vectors is minimal for a given p<w; (p=1 in fig. 1).

Typical results show that p<<w for a good approximation in a wide range of applications, including this one. Since the w PC's form a basis in R^W space, they can replicate exactly any of the n points in the set, using the scores as weights.

III Locating errors in a single strip

We will concentrate our efforts in locating errors in an elongated DTM defined over a regular grid of size w*n, w <<n. We will use the term strip to denote such DTM shape, and it is assumed that w corresponds to rows, and n to columns. Such DTM can be regarded as being composed of n cross sections (named profiles) each composed of sets of w points. The height can be referred to as h(i,j), being i bounded by w, and j by n.

In typical situations, strong correlation can be expected between "close" data points, both within each profile and between adjacent ones. On the other hand, isolated random errors are assumed to be weakly correlated with its neighbors. So a procedure may be designed to give a criteria for selecting a candidate h(i,j) as being an error, based upon some index or statistic that highlights a low correlation situation. Systematic errors might show a completely different behavior strongly related with the producing method, and they will not be considered here.

Some authors (Hawkins, 1974; López *et al.*, 1994; López *et al.*, 1993) attempted to locate errors in tabular datasets using PCA, but their results and methods could not directly be applied since DTM data cannot be automatically regarded as a time series nor a multiple replication of an experiment. Some different approach should be suggested.

A procedure with two steps is proposed. First, locate the profiles that are likely to hold an error, and secondly (within each of them) pick the location of the error itself.

a.1.- locate the columns (profiles) likely to have candidates

In order to highlight an unusual profile some properties of PCA will be exploited. They will be illustrated using the concepts already presented. What follows is based on the ideas presented in López *et al.*, 1994.

First question is how PCA is connected with errors. Fig. 1a shows the typical distribution of the score 1, for a single strip of the DTM test problem (which will be described later), and fig. 1b, for the second score. The distribution is build from n values of the scores derived for a DTM of size w*n (25*150 in this case). For this particular strip, the first 10 PC explain more than 98% of the total variance.

As it can be seen, the distribution does not show any special shape. The mean should be zero in all cases (figs. 1a to 1d). We want to emphasize that the first score distribution is not at all symmetric, which is more likely in the second one (fig. 1b).

The progressive evolution towards a symmetric distribution is clear in the fig. 1c and 1d, which illustrates the 20th. and 25th. scores distribution. Another property that should be noted is the decay in the "width" of the distribution as an index function. There is no profile for the fig. 1a and fig. 1b which first or second score has absolute value greater than 40 m. For the fig. 1c, the same property holds, with limits 0.8 m. for the 20th. score, and 0.9 m. for the 25th. one (fig. 1d). This decay is also related with the eigenvalues associated with each PC. Almost the same behavior can be observed for other strips. Notice that the scores share the units with the DTM data (m), but *they have sign*.

In all these figures there are two small arrows pointing to an "*" and to an "O". The first one points to the values of the scores for the profile presented in fig. 2a, and the other does the same for a slightly modified profile (fig. 2b) which has an isolated "error" pointed by the arrow which is not evident by looking at the profile alone, and will be left unnoticed in a 3-D like representation.

However, notice that for both the 20th. and 25th. scores (fig. 1c and 1d) the "O" value associated with the modified profile (fig. 2b) is now a marginal value, clearly separated from the rest of the samples.

This result holds in general for other strips. From the figures it can be seen that there are a few events (profiles, identified by the column) that renders marginally distributed values for *some* of the scores. What is the meaning of such events? Those events that behave in such a way are unusual events, and may contain the errors, as we have shown in the example. So the method will use some threshold interval for the j-score, and check all the events looking for the ones that lie outside the bounds (see Davies *et al.*, 1993). This is the first key idea.

The procedure (as sketched) disregard information coming from adjacent profiles. In other words, the profiles can be mixed, and the results will be the same. This information can in turn be used for handling some kinds of systematic errors, which is beyond our scope, or to extend the present method for non-isolated errors, which have not been considered at this stage.

Not all of the feasible set of scores will be used. Typically, the first ones are more related to physics (we meant by physics the underlying properties of the DTM) than to noise. That's why they are robust to single outliers (see the "O" in fig. 1a and 1b). So the first scores will not be checked. The appropriate limit for physics/noise qualification is subjected to direct experimentation, as will be presented later.

A related idea was proposed by Hawkins, 1974 who devised a similar approach. Instead of checking each score against a given (different) threshold, he computes the statistic T_2 and analyzes its distribution. T_2 is defined for profile (or sample) k as:

$$T_2(k) = \sum_{j=p}^{j=w} \frac{1}{\lambda_j} * a_j(k)^2 \qquad (2)$$

being λ_i the eigenvalues of the covariance matrix. In his work he says that the distribution of T_2 is a compound gamma function, assuming that the scores are normally distributed. Such property looks somewhat restrictive, but it is not strictly required. Any column profile is flagged if the value of T_2 is over a prescribed value.

Summing up, the approaches of both López and Hawkins take advantage of the fact that the non-systematic noise is likely to be important only in the weakest PC. The assumption that in this way it is possible to identify the columns containing the errors is important. What is remaining is how to identify the row for each candidate in any given column.

a.2 within each column (profile), find the rows that identify the candidates

T_2(k) is supposed to be "small" in most cases, except under the effect of the outliers, when at least one of the scores will be bigger than usual. Since each score is a linear combination of the height values of the column (due to the scalar product), the most-likely error row is chosen as the one whose variation mostly affects the value of T_2(k).

Therefore, a simple sensitivity analysis is carried out for each flagged "k" profile (column). The row that is responsible for the biggest change in the T_2(k) function, will be classified as an "a" candidate to be an error. Also "b" and "c" candidates are selected, in decreasing order of priorities. This is the other key idea.

There exists however some reasons for using a different statistics. We may create a positive function $T_2^*(k)$ (slightly different from the one of Hawkins, 1974) which is a weighted average of the squares of some of the scores a_j, for a given column "k".

$$T_2^*(k) = \sum_{j=p}^{j=w} W_j * a_j(k)^2 \quad (3)$$

Note that as before the summation starts on the p-th score. The weights W_j can be chosen to scale the scores so that all terms in the summation have the same order of magnitude. The scores can be related with the aforementioned eigenvalues, or fixed in another way. We have used for each j-score, a weight W_j that makes $\left| a_j(k) * W_j \right| \leq 1$ in 95% of the events. For example, if a_j is normally distributed with standard deviation σ_j, W_j will be exactly σ_j*1.95996.. (solution of $\frac{1}{2} erf(\frac{W_j}{\sigma_j \sqrt{2}}) = 0.95$). Hereinafter, 100-95%=5% will be called the penalty margin.

Such definition is more robust against the existence of outliers than the use of the eigenvalues. If the a_j are normally distributed, the statistic $T_2^*(k)$ is equal to the one of Hawkins except for a scale factor.

It should be pointed that $T_2^*(k)$ is not intended to be a weighted average of all the scores. Only those associated with the non-physical scores are included, according to the above mentioned limit p.

IV The proposed technique

Once we could locate errors in a strip, we could tackle the problem for a complete DTM. Given a regular gridded DTM model, with m rows and n columns, it can be viewed as formed by strips of width w rows and n columns. The union of all the strips is the whole DTM. In fig. 3 the whole matrix and a single strip is shown, while in fig. 4 the corresponding part of the DTM is presented.

For each row-wise strip, a set of "a", "b" and "c" error candidates can been selected, and its union will be called "row-wise candidates". This may be a complete solution for the problem. But why row-wise strips? There are no reason to discard analyzing the problem using column-wise strips. So doing the same for all column-wise strips will also cover the entire DTM and give three new global set of candidates. The union of the three sets can be called "column-wise candidates".

If the intersection of both sets ("row-wise candidates" and "column-wise candidates") is not empty, we have located those points in the DTM that behave atypically in both directions. These will form the *candidate set*, named also *guessed errors*.

The procedure involves five steps, and it can be sketched as:

Some remarks follows. We have used in the pseudo code a single value of strip width w both for rows and columns. This is not required but simplifies somewhat the tuning process, as will be presented later.

```
Given a DTM as a matrix of size m*n
subdivide the DTM in row-wise and column-wise strips of width w

repeat until criteria are satisfied:
        a) find row-wise candidate set:
                a.1.- locate the columns likely to have candidates
                a.2.- within each column, find the rows that identify
                        the candidates
        b) find column-wise candidate set:
                b.1.-locate the rows likely to have candidates
                b.2.- within each row, find the columns that identify
                        the candidates
        c) intersect both sets
        d) evaluate criteria
        e) correct all errors
end
```

The method is in fact iterative, since all the distribution functions (covariance matrix, etc.) are modified as soon as an error is removed. In each iteration, a candidate set is obtained with the presented methodology. After the candidates are corrected (if they are true errors) or they have been verified, they will be excluded from the feasible set of "a", "b" or "c" candidates, and a new iteration takes place. The process is supposed to stop when some criterion is fulfilled; for example, it stops if the type I error is too big (see below). Each iteration will be named "step" in the following discussion.

V The experiment:

To test the methodology we have used a Monte Carlo simulation procedure. Using a real DTM as a test problem we first selected at random about 5% of the points within the dataset. Then all of them were modified by adding an error, which were also randomly selected from a given feasible set, and finally the methodology is applied in order to locate the errors and its results were recorded for later statistical processing.

The DTM used as a test problem covers an area of 7.5x5 km in Stockholm with 150x100 points with a 50 m grid spacing and 1 m height resolution. The area consists mainly of hilly terrain, with height values ranging from 0 to 59 m. Fig. 5 shows a mesh view of the DTM, while in fig. 6 its height distribution is presented. The 0 m areas can be seen to the left of the fig. 5 . The DTM has a mean height value of 20.83 m and its standard deviation is 9.47 m. For this type of DTM, errors within [-4 m,+4 m] are typical. Since the data is rounded to the nearest meter, there will be little chance to pick errors of one meter. Any "feasible" error should also be an integer number.

There is scarce guidance in the literature about the spatial distribution of real errors, and they are strongly related with the generation procedure. Since we have a single replication of the DTM, we were not able to test the procedure with real errors. We follow some authors (Bethel *et al.*, 1984) in modeling errors as additive and isolated, being the added height chosen from a given set. As a feasible set we have used as a first example the values [-4,-3,-2,-1,+1,+2,+3,+4] meters, with equal probability, which is considered as a difficult case. This is expected to model spatially uncorrelated errors, and we named it as *spike-like* errors (see fig. 7, left).

The selection of 5% as a typical value looks somewhat high when compared to the one reported by Östman, 1987. He found a typical value of 0.5% for the number of *gross* error occurrences, but here the worst errors are of absolute size 4m and they account for only 1/4 of the total.

As another alternative for an error shape model, we also tried a more structured one, which resembles a pyramid; once a point is selected, it is modified by adding a 2Δ meter error, and the eight points surrounding it adds only Δ (see fig. 7, right) We have selected Δ uniformly from the set [-2,-1,+1,+2]. We named this model *pyramid-like*, and it is expected to model some degree of spatial correlation in errors.

Once the model has been contaminated with errors, the described procedure is applied. Since one possible application for the method is to help locate errors while the model is being created, various measurements of success have been provided. The most important is the probability of type I error (Ounping, 1988), which measures the probability of classifying a correct value as wrong. It is estimated with the quotient between the number of good points classified as errors, in relation to the number of candidates suggested. This statistic can be calculated for each step (even in real applications), and the user can take the appropriate decision to continue the process or to stop.

Also, the relative importance of the errors is important, and not merely how many they are. Since in this test we know in advance how the original errors are distributed, the distribution of the identified errors can also be calculated. That will not be possible in the real operation, but it will render useful information at this stage. We will show results for errors of any size, but most figures will concentrate on errors of absolute size 4 m. The reason will be clear later. The number of errors remaining in the dataset (classified as good by previous steps) in relation with the initial population evolves monotonically as the process continues but might become stationary (no more errors are found irrespective of the number of iterations).

The method has some free parameters, and some previous estimations should be done in order to fix its value. They include

a) which scores will be considered as associated with noise.

b) what threshold interval will be used, to separate the marginal from the typical values in each distribution.

c) the penalty margin value

The strip width has been kept fixed, that means w has been held constant for all strips (both row-wise and column-wise). Instead of selecting a different number of scores for each strip, a single value has been chosen.

Further theoretical guidance for the choice of w for a given m and n has not been investigated in full. The number w can be assumed as a common divisor of m and n, and the minimum ratio n/w can be derived following Hawkins (1974). He gives some theoretical results, assuming that all the scores are normally distributed. Figure 8 gives guideline values depending on the confidence value α, and they are derived after Hawkins, 1974, 1994. We used n/w=150/25, 150/15 and 150/10 at best in our examples, which are denoted in the figure as "+", "o" and "*" respectively. They imply somewhat low values for the confidence α. Under some assumptions (Hawkins, 1974) these results may indicate that our results using w=10 should be more reliable than those of w=25.

However, from our results will be clear that given a DTM an appropriate value for w can be estimated by means of a similar experiment that the one in this paper. Further guidance will be given later.

Different and separate series of experiments were carried out, using values for w of 10, 15 and 25 elements and also applying *spike-like* and *pyramid-like* shaped errors.

From some runs that were done previously, it became clear that once w was fixed, the most important parameter is the number of uncontrolled scores, being the limit between the physics and the noise. This limit may be found in each particular DTM by means of a simulation like the one described in IV, or by applying some rule of thumb.

The other parameters were initially fixed at 2.5% for the threshold level, and at 5% for the penalty margin. Increasing the first one renders more candidates to the "a", "b" and "c" sets by means of selecting more columns (profiles). In order to avoid an empty candidate set, we increased the threshold level if too few candidates are suggested. The threshold interval is derived from the cumulative sampled histogram, using appropriate bounds. The penalty margin, provided it is not too small, is somewhat insensitive to the gross-errors, and so are the weights W_j.

The goal is to minimize both the probability of type I and II error. However, not both of the objectives are equivalent, and in some situations one may be more important than the other, depending upon the user. This will be discussed later.

VI The results from the Monte Carlo simulation

Results for spike-like errors:

We will denote as *candidates* or *guessed errors* the group of coordinates (i,j) suggested by any single step of the procedure. The *true errors* are those elements in the previous set that also belong to the known errors set.

Fig. 9a shows the average Type I error evolution up to 5 per cent depuration effort (see below). The y-axis shows the evolution of the type I error calculated as the number of points missclasiffied as errors compared with the number of candidates, averaged after 50 replications of the random error set. The x-axis shows the effort, defined as the fraction of the dataset already revised. An effort of 100 per cent implies that all possible points have been checked, since the effort per step depends on the number of uncontrolled scores and the threshold level. We interpolated our results to prescribed effort values using splines. Each polyline corresponds to different strip width and also number of uncontrolled scores. For w=10, 15 and 25 we left uncontrolled 6 scores.

From this result it is clear that in the first 1 per cent effort the measured Type I error is low, being below 5 per cent for the lower values of w. The dashed horizontal line corresponds to the limit of 95%. Obtaining an error rate over that level is worse than to pick the points at random, since that level is the noise initially seeded into the DTM. The limit was shown as being constant, despite the fact that such probability grows slightly as soon as an increasing number of errors has been found. For 2 per cent effort and over somewhat poorer results may be achieved, but certainly better than chance.

But good results in the type I error is not the whole picture. From a DTM producer's point of view, what is more important is to minimize errors still in the DTM, so the type II error is more representative. It measures the probability of classifying as good a wrong value. The type I error in fig. 9a only *count* the success and the failures, but do not reflect the relative importance of the errors. The absolute value of the error is not considered.

Table 1 shows the evolution for a single replication of the experiment of the distribution of the errors. The original distribution of the error size is shown in bold. For example there were originally 116 points with error +3. The rows below show the remaining errors after finishing each step of the cleaning process. For example, after step 4 only 49 of the 116 original errors of size +3 remain in the dataset.

As expected the errors of size +1 and -1 were very difficult to locate, since the original DTM has a resolution of that size. We omitted the corresponding effort involved.

Size of errors	-4	-3	-2	-1	1	2	3	4	Total
Original # errors	**93**	**85**	**89**	**82**	**83**	**98**	**116**	**93**	**739**
# after step 1	36	53	78	82	81	83	72	43	528
# after step 2	21	26	66	82	79	74	56	29	433
# after step 3	20	19	57	81	79	69	52	25	402
# after step 4	18	17	50	80	78	68	49	24	384
# after step 5	18	17	48	80	78	67	47	24	379

Table 1 Error´s size distribution for a single experiment, when 5 terms are left uncontrolled

We limit our Type II error calculations to errors of absolute size exactly 4m, because as table 1 shows, the other cases are less prone to be located and its type II error will be unaffected. In fig. 9b the evolution of the Type II error is presented. Notice that the best results are for w=25, but for w=15 very similar results are achieved. The initial Type II errors is 1.21% in all cases, so it can be reduced to 0.64 with only 1 per cent effort.

The same behaviour were noticed for other combinations of w and number of uncontrolled scores; better results are obtained for the Type I error for lower w values, while the opposite happends for the Type II error.

In fig. 10 the type I error up to the 1 per cent effort is represented as a function of the number of scores not controlled. The continuos line were obtained by spline interpolation. Notice that, irrespective of w, the relative minimum lies between 5 to 10, and the absolute minimum correspond to the option w=15, nearly twice the optimum number of uncontrolled terms.

In fig. 11 the Type II error up to the 1 per cent effort is analyzed in relation with the number of uncontrolled scores. The results for w=10 are worse than the other options, but it is not so evident the differences between w=15 and w=25. So a bigger w is preferred, while the optimum choice for the scores to be left uncontrolled is less crucial. On the other hand, from fig. 10, the w=10 option is worse as far as the type I error is concerned, while w=10 results in similar figures.

This result supports some conclusions:

a) the optimum choice for the number of controlled scores is more or less independent of the goal (minimize the type I or II error), while the situation for w is slightly different. In the former, a smaller w is preferred while in the later a bigger one is better.

b) there seems to be a compromise w value for a given DTM, which can be estimated from simulations before truly apply the methodology, or in another way. In this case it should be near w=15.

c) the results for type II error are based only on errors of size 4m. Slightly different figures can result from considering other categories.

As mentioned before, a crude estimation for the optimum w can be done without performing a simulation. From the outlined results, the best w is nearly twice the optimum number of uncontrolled scores. That value should be in turn near the limit between the physically meaningfull and the noisy eigenvectors, and to define the former there is no unique rule in the literature. We applied the one suggested by Hawkins, 1974 which in our case proves to give similar values for different w. See the original reference for a justification. The procedure might be the following:

1) choose some suitable initial value, based on fig. 8.

2) for each strip, calculate the covariance matrix, and its eigenvectors. We will assume that they are sorted in ascending order, depending on the corresponding eigenvalue. Find the p that makes

$$\min_{j=1..p}\left(\max_{i=1..w}\left(abs\left(e_{ij}^{(p)}\right)\right)\right) > 0.25 / \sqrt{w}$$

Since this value depends on each strip, we select the median of the results for all strips. In our DTM, for w=10, 15, 20, 25 and 30, we obtain p=3, 4, 4, 3 and 4. Estimate the optimum number of uncontrolled scores as twice that number and the proper w as four times. In our case, it will suggest 6 to 8 uncontrolled scores, and for w something between 12 and 16.

3) choose the definitive w as close as possible to the value obtained in the previous step, while being a divisor for both m and n.

This will give a rough estimate, while a simulation like the one described in this work will give more reliable results. Simple or flat terrain will be well represented with less terms, so w will be smaller, while for complex ones it should be greater.

Results for pyramid-like errors

In previous results, all the calculations have been performed assuming that typical errors are completely isolated ones, uncorrelated in space. This is not true in practice even though the shape of the errors has not been analyzed in the literature. Bethel *et al.*, 1984 used spike like errors only, and we present here the results of a different model: the pyramid-like error.

Since pyramid errors are spatially autocorrelated, the chance of locating them is lower. That´s due to the underlying design of the methodology, strongly oriented towards independent errors in space.

There are also some other details regarding the "accounting" procedure. We will consider as a candidate not only any point which is both a row-wise and column-wise candidate, but also its immediate neighbors. So for every candidate, nine points are checked. However, due to computational simplicity no effort has been made to take into account the overlap of candidates for the same step (i.e. if both a point and its neighbor are selected, there are points that count twice). So the results are somewhat pessimistic in terms of the type I error.

The behavior observed in figs. 12a and b are very similar to the ones observed for the spike-like error shape model, althoug the numbers are more pessimistic. The first 1 per cent effort renders an acceptable Type I error, but the second and others are somewhat higher. The analysis of the Type II error should take into account that the initial value in this case is 0.81%, while in fig. 9b it was 1.21. This imply that the procedure reduces at most the Type II error in 79% with only 1 per cent effort.

For comparison purposes, fig. 13 shows the type I errors for the first step in terms of the uncontrolled scores, for different values of w, and for 50 replications of the experiment.

The type II values in fig. 14 shows that (as happened before with isolated errors) the weak dependence upon the number of uncontrolled scores; the associated error is fairly "high" (0.64% per cent up to the 1 per cent effort, and 0.42 up to the 5 per cent effort, starting with an initial value of 0.81%,) when compared with the other error shape model. This imply that 79% of the gross errors remain in the dataset with 1 per cent effort, and 51% cannot be located even with a 5 per cent effort. So again, as expected this results are worse than those for isolated errors, but still might be of use.

Summing up, as expected for the pyramid error shape, for most of the cases the 1 per cent effort still renders a low type I error, being rather insensitive to the number of uncontrolled terms (but in the range 4 to 8) and to the alternative w=10 or w=15.

In terms of the type II error, about 21 per cent of the worst errors can certainly be located with only 1 per cent effort of the procedure, and it may go up to 49% with 5 per cent effort. The best results follow the ones obtained with isolated errors, being w=25 the best option. The number of uncontrolled scores is again between 5 and 10.

The conclusion is that the method proves to be effective in identifying a significant amount (up to one third) of the big errors with limited effort. For better Type I results, a smaller w is suggested, while for Type II optimization a somewhat greater might be the option irrespective of the shape model assumed. The number of uncontrolled scores is between 5 and 10 in any case.

No special pattern of the location of the errors found were noticed during the runs. Such aspect may be investigated in the future, with a wider set of DTM.

VII Discussion:

From the results obtained, a process can be devised to pinpoint an important part of the larger random errors in a raster dataset. Further actions strongly depend on the application the user is involved with.

In a production environment, some action can be taken to check these identified isolated values. In photogrammetric measurements these checks can be done before removing the stereopair. The goal here is to remove most of the errors, i.e. diminishing the type II error, while the type I error is less crucial.

On the other hand, the end user is left alone in most cases, because he may not be able to go to the original data sources. Therefore he should be worried by the risk of modifying a value that is correct, so the type I error is more important.

The results show that it can be assumed that up to the 1 per cent effort, most candidates are errors. The associated Type I error can be less than 5%, as has been shown for isolated errors, and around 25% for pyramid-like ones for proper choice of the parameters. The Type II error is defined here only for errors of absolute size 4 m, and it can be reduced 64% (for isolated errors) and 21% (for pyramid-like errors) checking only 1 per cent of the dataset.

Every step produces a candidate set, and once this set is obtained, any standard procedure can be used to replace the outliers with suitable values. As long as the dataset is progressively being corrected the risk that a point classified as an error is correct is higher, and some caution should be taken.

The procedure may be unable to locate non random errors, i.e., if a region has been affected by an improper choice of the control points, or there are edges along the rows or columns that arise from an improper matching of a partial DTM, for example. Further studies will be necessary to clarify this aspect.

The test area is considered to be a difficult one. Rough terrain, narrow channels, steep hills, and small water areas are typical, all of them may easily mask errors. The DTM itself should not be considered as free of errors, and it has been used "as is". This fact is common to most users of this kind of data, so it is believed that such a situation will not limit the range of applications of the ideas presented.

It should be noticed that there are two integer parameters free: the strip width and the optimum number of scores that should be left uncontrolled. Here we have taken a fixed strip width w for both row-wise and column-wise strips, and we have considered a single value also for the optimum number of scores for all strips.

Loosely speaking, the optimum number of uncontrolled scores should be somewhat stable, provided that w is not too small. This assertion comes from a weather analogy. In that case, increasing w is the same as adding a new weather station to the set. The optimum value is related with the number of typical weather systems, and that is certainly independent of the number of observations being taken.

If w is less than this (unknown) optimum value, poor results will be achieved. We noticed it using w=10. On the other hand, if w is too big, the number of events will be small compared with w, and unstable results may appear. In the extreme case of w>number of events, the covariance matrix will be no longer positive definite, and zero eigenvalues will appear. This optimum may serve as an objective number which characterize terrain complexity being lower for smooth terrain. Further work with other DTM may render a closer relationship between roughness and this number, also linked with grid size. We provide some rough estimation rule, which has to be tested in other cases.

There are two other non-integer parameters: the penalty margin and the threshold level to be considered. Some calculations have been carried out and the penalty margin value has been found to have a rather low influence on the final result.

The threshold level is the key for the amount of work to be done. If its value is too low, only very few values will be chosen as candidate errors. They certainly have a good chance to be true errors (i.e., the Type I error is lower). But some others, which are also errors, may not be picked even in further steps (i.e., the Type II error will stay high). This should be the choice for an end user.

On the other hand, if its value is higher, more candidate errors will be selected, but the efficiency will be lower. That also implies that the type I error may be higher, and that may be unacceptable. In an automatic production environment, where there is a chance to check the values, this may not be a problem. In a semi-automatic one, the operator may quickly become bored of checking points that are correct, and the overall procedure may be considered as poor, even if most of the errors are in fact removed.

The end-user has limited ways to check the values. Maybe he can use a three-dimensional plot of the interpolated surface near the candidates, or a contour plot, or something else to get an idea of the surroundings. But taking into account that a) he certainly will not check that way too many candidates, and b) he also has no definitive way to find the true value, he will limit the search to a small set, where may be included the "worst" errors. So he will choose a somewhat lower threshold.

A comment about the computer time requirement: the procedure involves for each step, the computation of (m/w).(n/w) covariance matrices of size (w,w), which takes $\left[(n/w).O(n^2)+(m/w).O(m^2)\right].O(w^2)$ operations, find its eigenvectors after $\left[(n/w)+(m/w)\right].O(w^2)$ operations, and project each strip to calculate the scores (which in turn requires (m+n).w operations). Some other operations are required but depend linearly on m and n. In our example, for a DTM of size m=150, n=100, and for w=10, it requires about 3 seconds in a PC486 (and 2 in SUN Sparc 10) per step, both working with MATLAB, so the overall procedure is considered cheap in terms of computer time.

VIII Conclusions:

A new methodology to locate random errors in quantitative raster data has been presented, and tested in a grid-based DTM as an example. The methodology is iterative, and proves to be robust, rendering type I error rates near 5% for isolated errors. Roughly speaking, it also located half of the 4 m isolated errors checking only 1 per cent of the database.

The process involves the decomposition of the DTM into strips, and requires a Principal Component Analysis (PCA) of each one. That is not the usual way of using the technique in image processing. Further simple calculation renders three sets of candidates to be considered. The stripping process is done both row-wise and column-wise as a cross check, and a even more reduced set of candidates is obtained.

Some experiments were performed using a DTM with heights between 0 and 59 m seeded with randomly located additive errors, with amounts up to 4 m. This value was considered to be on the order of the errors in the model. Two error shape models were considered: one completely isolated (like a spike) and the other with some arbitrary regular shape (pyramid like). Even though it has been assumed that those shapes are typical, their representativeness for real DTM errors is still to be investigated.

The method has some parameters left free to the user, and some guidance is provided. However, at least for the first candidates, the high rate of success obtained proved to be fairly insensitive to some of the parameters.

In the case of using the algorithm in a semi-automatic production environment, the method warns the operator about possible errors before the stereopair is unmounted, enabling a new measurement. In a fully digital production environment, some correlation thresholds have been usually fixed weakly to diminish computer time. This method may help in selectively strengthening the correlation thresholds in unlikely points.

In the case that there is no possibility to verify the errors, e.g. for end users, the algorithm will help to locate the most unlikely values; they may be replaced with the aid of some suitable interpolation method. If there are some independent sources (cartographic maps, etc.) they could be used for checking.

IX Acknowledgments:

This work was started during a stay at the Environmental and Natural Resources Information Systems of the Royal Institute of Technology, Stockholm, Sweden, under the supervision of Prof. Friedrich Quiel. Dr. Juan Echague made several suggestions which certainly improved this work. The author wishes to acknowledge BITS (Swedish Agency for International Technical and Economical Cooperation) and CONICYT (Consejo Nacional de Investigaciones Científicas y Técnicas, Uruguay) for partial funding. Some of the basic routines have been programmed by Elías Kaplan, and his collaboration is also acknowledged. Computer resources have also been kindly supplied by the Department of Numerical Analysis and Computer Science at the Royal Institute of Technology, and by the Centro de Cálculo of the School of Engineering, University of the Republic, Montevideo, Uruguay.

X References:

Ackermann, F. 1995. Digitale Photogrammetrie - Ein Paradigma-Sprung. Zeitschrift für Photogrammetrie und Fernerkundung, 3/95, 106-115

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

Bethel, J. S.; Mikhail, E. M. 1984. Terrain surface approximation and on-line quality assessment. International Archives of Photogrammetry and Remote Sensing. Commission III, V25, A3a, 23-32

Chavez, P.S.; Yaw Kwarteng, A. 1989. Extracting Spectral Constrast in Landsat Thematic Mapper Image Data using Selective Principal Component Analysis. Photogrammetric Engineering and Remote Sensing, V 55, N 3, March 1989, 339-348.

Davies, L.; Gather, U. 1993 "The identification of multiple outliers" Journal of the American Statistical Society. Sept. 1993. V 88, N 423, 782-801.

Day, T.; Muller, J.P. 1988. Quality assessment of Digital Elevation Models produced by automatic stereo matchers from SPOT image pairs. International Archives of Photogrammetry and Remote Sensing, V27, B3, Commission III, 148-159.

Eklundh, L; Singh, A. 1993. A comparative analysis of standardized and unstandardized Principal Components Analysis in remote sensing. Int. J. Remote Sensing, V 14, N 7, 1359-1370.

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

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

Hawkins, D. M. 1994. Personal communication

John, S. A. 1993. Data integration in a GIS - The question of data quality ASLIB Proceedings, V 45, N 4, 109-119.

Lebart, L.; Morineau, A.; Tabard, N. 1977. Techniques de la description statistique: Methodes et logiciels pour l'analyse des grands tableaux. Ed. Dunod, Paris, 344 pp.

López, C.; Kaplan, E. 1993. "Principal Component analysis applied for outlier location and missing value problem wish surface wind and pressure data" Internal report, 22 pp. Available from http://www.fing.edu.uy/~carlos/papers/rep93_5/viento.htm.

López, C.; González, E.; Goyret, J. 1994. Análisis por componentes principales de datos pluviométricos. a) Aplicación a la detección de datos anómalos. Estadística V46, N 146-147, 25-54.

Östman, A. 1987. Quality control of Photogrammetrically sampled Digital Elevation Model. Photogrammetric Record, 12 (69) 333-341.

Ounping, Gong. 1988. Experiment and discussion on the method of blunder location. International Archives of Photogrammetry and Remote Sensing, V27, B3, Commission III, 841-849.

Richman, M. B. 1986 Review article: Rotation of principal components. Journal of Climatology, V 6, 293-335.

Theodossiou, E. I.; Dowman, I. J. 1990. Heighting accuracy of SPOT. Photogrammetric Engineering & Remote Sensing, V 56, 1643-1649.

Figures

Appendix 5

López, C., 1997, "On the improving of height accuracy of Digital Elevation Models: a comparison of some error detection procedures" *Sixth Scandinavian Geographic Information Systems Conference, Stockholm, 1-3 June, 85-106*

PROCEEDINGS ScanGIS 1997 85-106

On the improving of elevation accuracy of Digital Elevation Models: a comparison of some error detection procedures

CARLOS LÓPEZ
Centro de Cálculo, Facultad de Ingeniería (11), Universidad de la República
Julio Herrera y Reissig 565, Montevideo, URUGUAY
internet:carlos@fing.edu.uy, http://www.fing.edu.uy/~carlos

Abstract: The widespread availability of powerful desktop computers, easy-to-use software tools and geographic datasets have raised the quality problem of input data to be a crucial one. Even though accuracy has been a concern in every serious application, there are no general tools for its improvement. Some particular ones exist however, and we are reporting here results for a particular case of quantitative raster data: Digital Elevation Models (DEM). We tested two procedures designed to detect anomalous values (also named *gross errors*, *outliers* or *blunders*) in DEM, but valid also for other quantitative raster datasets.

A DEM with elevations varying from 181 to 1044 m derived from SPOT data has been used as a contaminated sample, while a manually derived DEM obtained from aerial photogrammetry has been regarded as the ground truth. That allows a direct performance comparison for the methods with real errors.

We assumed that once an outlier location is suggested, a "better" value can be measured or obtained through some methodology. The options are different depending upon the user (end users might only interpolate, while DEM producers might go to the original data and make another reading). In this experiment we simply put the ground truth value.

Preliminary results show that for the available dataset, the accuracy might be improved to some extent with very little effort. Effort is defined here as the percentage of points suggested by de methodology in relation with its total number: thus 100 per cent effort implies that all points have been checked.

The method proposed by López (1997) gave poor results, because it has been designed for errors with low spatial correlation (which is not the case here). A modified version has been designed and compared also against the method suggested by Felicísimo (1994).

The three procedures can be applied both for error detection during the DEM generation and by end users, and they might be of use for other quantitative raster data. The choice of the best methodology is different depending on the effort involved.

KEYWORDS: DEM, quality control, blunder location, gross error location, accuracy

1. Introduction

Geographic Information Systems (GIS) is one of the fastest growing markets in software today (Anon 1994). That implies that more people have access to proper tools, and then are able to manipulate and produce data. Data availability will be assured in the future, through the operation of the so called *Clearinghouses*, which will distribute existing datasets to government, industry and the general public (Nebert 1995, 1996).

The combination of widespread data and ready made, easy to use software raises some critical points. John (1993) stated that "...very wrong answers can be derived using perfectly logical GIS analysis techniques, if the users are not aware of the particular peculiarities of data...". Data quality is emerging as one of the most important issues in GIS technology for the next years. Its management requires methods to describe, visualize and measure it properly (see Hunter *et al.* 1996). Standards for describe the quality are presently under development.

Thapa *et al.* (1992) remarked that when setting up a GIS, most of the costs (maybe up to 80 per cent) are related to acquiring and/or collect data. Once the dataset is obtained further efforts to improve accuracy should be as effective as possible. This paper reports some results on that subject. We will concentrate here on Digital Elevation Models (DEM). We will not consider errors in the intermediate steps in the process of DEM generation, but we will concentrate on the errors in the final product. According to Thapa *et al.* (1992) errors can be classified into three types: (1) gross errors and blunders, (2) systematic errors and (3) random errors. Gross errors and blunders are caused by carelessness or inattention of the observer in using equipment, reading scales or writing down readings, etc. They can also be caused by malfunctioning of the equipment. Observations affected by this kind of errors are useless, and should be eliminated. From a statistical point of view they cannot be considered as belonging to the same population as the other observations. Systematic errors occur in accordance with some deterministic system which, if known, may be represented by some functional relationship. In a statistical sense, systematic errors introduce bias in the observations. Unlike gross errors, they cannot be detected or eliminated by repeated observations (the errors may be *precise*, but they will not be *accurate*). After removal of gross and systematic errors, differences still exist due to random errors. They cannot be removed by repeated observation, and they cannot be modeled with a deterministic relationship. If sufficient observations are taken, random errors posses the following characteristics: a) positive and negative errors occur with almost the same frequency b) small errors occur more often than large errors and c) large errors rarely occur.

Östman (1987a) pointed out the fact that there exists no unique criteria or single measure for the "quality" of a DEM. He suggested that one should at least, consider accuracy in elevation, slope and also curvature. However, accuracy reports in terms of slope are very unusual. An exception can be found in the work of Giles *et al.* (1996) who compared a 20 m resolution DEM derived from SPOT images with field measurements in terms of slope. They recognized that the elevation error might have two components at different scales. To filter out the small scale error they simply applied a 3 by 3 median filter and to remove the larger errors they used a 11 by 11 window, with a different filter. They claim that such filtering improve to some extent the accuracy in slope, without significantly degrading the accuracy in elevation. This has also been reported by Östman (1987a). He found that the RMS error in elevation decreases with decreasing grid size (as expected) but the effect in RMS error in slope is very limited. Förstner (1983) gave theoretical arguments for this fact.

Accuracy of photogrammetrically sampled DEM depends on the data sources and the procedures involved. It has been a considered an important problem which led to collective efforts like the one summarized by Torlegård *et al.* (1986). They reported the results of DEM derived independently by a number of organizations working on the same set of aerial photographs. Six different terrain types have been chosen, ranging from smooth terrain to steep and rugged mountains. They found that the errors of the elevations in photogrammetrically measured DEM consist to a large extent of systematic components. Regarding error location, they applied a "rule of thumb" based on recursive filtering using a 5 by 5 window, and declared that everything located is an error. They conclude that the number of those so defined errors typically varies between 0 and 3 per cent. They noticed that error size is independent of terrain type and that errors are more frequent in difficult terrain.

A similar (deterministic) approach was used in an early paper by Hannah (1981), who detects errors by applying constraints to the slopes and to the changes in slope at each point. Felicísimo (1994) analyzed the differences between the elevation and an interpolated value from the neighbors. Assuming gaussian distribution of the errors, he analyzed the differences by means of a standard Student t test.

Using the Torlegård *et al.* (1986) dataset, Li (1992) analyzed the dependence of the final accuracy on the sampling interval. His starting point is the gridded data and he degrades it by subsampling. He used several measures of accuracy: the RMSE (root mean square error), the mean μ, standard deviation σ and maxima of the difference between "truth" and data. "Truth" is available at selected checkpoints derived from larger scale photography. He found positive correlation among the RMS error in elevation and the slope of the terrain.

Day *et al.* (1988) tested three methods for the generation of DEM based on SPOT data. The three results were compared with a very carefully, manually digitized 30 m grid DEM, in terms of elevation differences. Even though the goal of the work was to compare the operational behavior of the algorithms, they do not propose a solution for the location of the errors. The distribution function of the absolute size of such errors is also presented for each method. They also reported how many checkpoints lie outside the limit $|\text{error}-\mu|>3\sigma$.

Any method for locating the errors should make assumptions about size, location and spatial self correlation. Bethel *et al.* (1984) proposed the method of maximum chi-squared ratio for on line quality control, and tested the methodology using uncorrelated in space, spike-like blunders of no more than 10 feet (about 3 m). López (1997) used two error models: one uncorrelated in space (spike-like blunders) and another weakly correlated (pyramid-like).

In the field of Image Processing the term *salt-and-pepper* has been coined for weakly self correlated errors. They are routinely corrected using filters. The most popular and simple one is the median filter (Mitra *et al.* 1994) but it has the fundamental inconvenience that it smoothes out all the DEM; current efforts are directed towards a division of the problem: to separate error detection from error correction, and to use state variables for error detection (Abreu *et al.* 1996).

We will not discuss here the methods for obtaining the DEM itself. There are well established procedures based on photogrammetry, GPS, etc. However, if the equipment or the methods are at their limit today, there will be little chances to improve the final results by merely pointing out some locations likely to be in error. Fortunately, this is not the case. Ackermann (1995) points out that the trend in DEM production is towards a move from interpolation to approximation, because the new generation equipment is able to produce many elevation values, but possibly with less accuracy than traditional equipment. The surface is *approximated* using many points, instead of being *interpolated* from few, very carefully obtained values.

Summing up, accuracy is a concern for the data producer as well as for the end user. Accuracy is usually described using different statistics of the distribution of elevation error at some checkpoints.

This paper presents test results of some recently proposed methodologies for locating errors which can be applied both by the producer and the end user. The methods were tested in a DEM with real errors, and the results are presented. Also some guidelines for the error model for this particular case are presented.

The paper is organized in eight sections. Section 2 has a brief outline of both the already existing techniques and the modified technique. Section 3 describes the data and summarizes some of its statistics. In section 4 the performance of the three methods is compared for the test DEM. Finally, section 5 contains a discussion and section 6 is devoted to conclusions, where the results and proposed future work are discussed. Acknowledgments are included under headings 7 and References appear at the end.

2. The error detection procedures in brief

For the sake of completeness we will describe briefly the methods of Felicísimo (1994) and López (1997), and a modification of the latter.

2.1 *The method of Felicísimo (1994)*

This is the simplest method available for this problem. Assuming that outliers are only locally correlated, the method analyzes the differences $\delta_{i,j}$ between the elevation value $z_{i,j}$ and an interpolated guess $\hat{z}_{i,j}$ obtained from its immediate neighbors. Assuming that the difference has a Gaussian distribution with mean $\overline{\delta}$ and standard deviation s_δ (both obtained from the sample) a Student's t test can be applied to validate the hypothesis that $\delta_{i,j}$ belongs to the population of deviations. Operationally, we analyze the statistics $t_{i,j} = \left(\delta_{i,j} - \overline{\delta}\right) / s_\delta$ which can be regarded as a standardized deviation. Since the number of data points are usually large, we can assume a distribution $t_{\alpha[\infty]}$ for $t_{i,j}$. For α=0.001, the statistical $t_{\alpha[\infty]}$ has a value of 3.219 for a two-tail test, where the null hypothesis is $H_0, \delta_{i,j} = \overline{\delta}$ and the alternative is $H_1, \delta_{i,j} \neq \overline{\delta}$.

We used a best fit approximation with a biquadratic polynomial using the eight closest neighbors to calculate $\hat{z}_{i,j}$. Along the borders we assume a mirror symmetry, and in the corners we used a linear interpolation with the three closest values available. We point out as candidate to be in error any $\delta_{i,j}$ that makes. $|t_{i,j}|$>3.219. The author states that even though a significantly high value of $t_{i,j}$ does not necessarily imply an error, it is an excellent alarm sign. We will analyze this topic later.

Once an error is located and corrected, both statistics $\overline{\delta}$ and s_{δ} change and new candidates appear. The method can be iterated and it might stop if no more "outlying" values remains. This is undesirable because we know that there still are errors in the dataset, so we proceed by lowering the limit 3.219 to 3.0 at least once. The new candidates once corrected modify the statistics, and new candidates with the limit 3.219 appear.

The method appears to be extremely simple and is parameter free. In section 3 we will investigate if the test DEM fulfills the assumptions under which the Felicísimo's method can be applied (Gaussian distribution, etc.). Also the relationship of $t_{i,j}$ and real errors (available in this experiment) will be presented.

2.2 *The method suggested by López (1997)*

The author proposed that any given raster dataset can be analyzed by means of subdividing it into elongated strips (figure 1). Each strip is assumed to have length n and width w (w<<n). The method considers the strip as a set of points in the R^W space. Each cross-section is represented by a point, where the original elevation values establish the w coordinates. The case of w=3 is illustrated in the figure 2, where each point M_k represents a cross-section.

The error location procedure directly analyzes the cloud of points in R^W, disregarding any order among points. This is an important assumption, since the concept of spatial self correlation looses completely all significance in the cloud. Adjacent profiles (of length n) need not to be in any special order, since they are coordinate axes in the space R^W.

The use of the cloud is common practice in statistics (Hadi 1992, 1994, Hawkins 1974, 1993a, 1993b), since the notion of "spatial correlation" and "precedence" is meaningless in most tabular data.

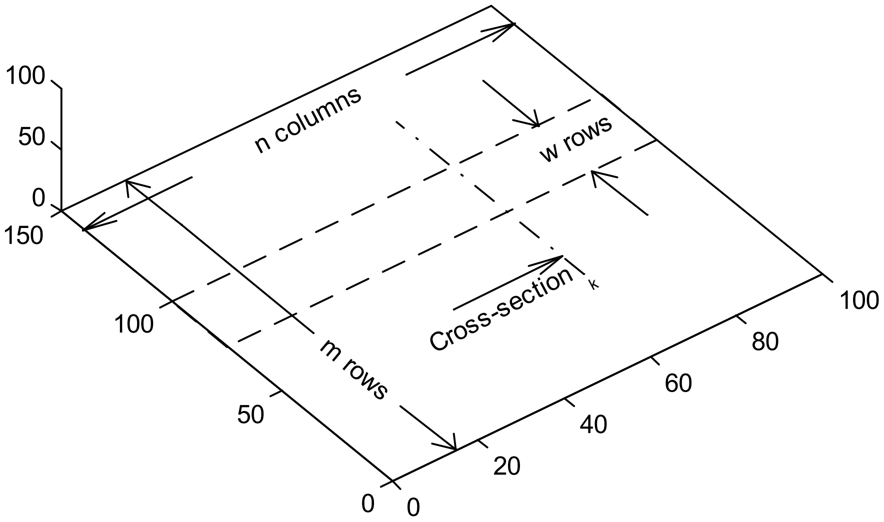

Figure 1 Sketch of the strip notation

The procedure is based upon Principal Component Analysis (PCA), which attempts to find the direction $\mathbf{e}_1$ of the vector in R^W space which minimizes S, defined as the sum of distances M_k-H_k squared taken over all k (figure 2). The origin **O** is the centroid of the set of points. For the sake of clarity, points with negative coordinates are not shown in the figure.

The projection O H_k, which is also the scalar product of vector $\mathbf{M}_k$-$\mathbf{O}$ with the unitary vector $\mathbf{e}_1$, is called the score (after Richman 1986). Thus $\mathbf{M}_k$-$\mathbf{H}_k$ is orthogonal to $\mathbf{e}_1$. There is one score value associated with vector $\mathbf{e}_1$ for each point in R^W. Let us also assume that $\mathbf{e}_1$ is unique.

If all the values M_kH_k are zero, we have reduced the problem of original dimension w, to a one-dimensional one. All the variability in the observations is explained by a single vector $\mathbf{e}_1$. If this is not the case, we may try to repeat the procedure with the remaining variability M_kH_k, which belongs to a (w-1) subspace of R^W orthogonal to $\mathbf{e}_1$. The original measurements $\mathbf{M}_k$ - **O** can be replaced with the difference $\mathbf{OM}_k$ - $\mathbf{OH}_k$, which is equal to $\mathbf{M}_k$- $\mathbf{H}_k$.

For the new cloud there should be a vector $\mathbf{e}_2$ (orthogonal to $\mathbf{e}_1$) which minimizes the distance S in the R^W space. The process continues until w vectors $\mathbf{e}_p$ have been created; each new vector $\mathbf{e}_p$ being orthogonal to all the previous ones. The vectors $\mathbf{e}_p$ are called principal components (PC).

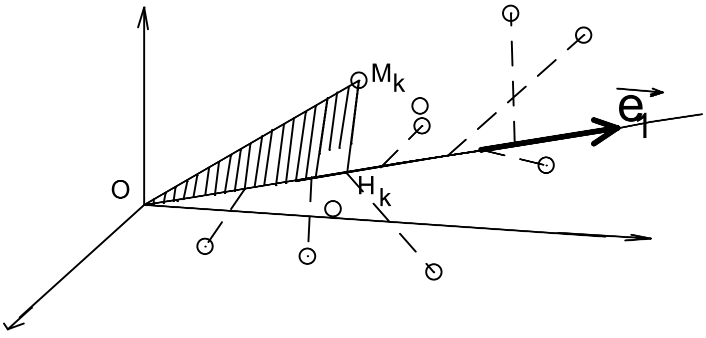

Figure 2 Sketch of the first principal component, for w=3

Each event $\mathbf{M}_k$ - **O** can be expressed as a linear combination of the PC's

$$\mathbf{M}_k - \mathbf{O} = a_1(k) * \mathbf{e}_1 + a_2(k) * \mathbf{e}_2 + a_3(k) * \mathbf{e}_3 + ... + a_w(k) * \mathbf{e}_w \quad (1)$$

It can be shown that the scores a_i associated with vector $\mathbf{e}_i$ are uncorrelated with those of vector $\mathbf{e}_j$. The vectors $\mathbf{e}_i$ are the eigenvectors of the covariance matrix of the data, and its components are named *loadings* in the literature. The sum of the corresponding eigenvalues equals the sum of the squares of the distances M_kH_k (Lebart *et al.* 1987).

PCA analysis generates a sequence of principal components, which explains most (or all, for p=w) of the variance of the data. This implies that the RMS error in approximating the data with a linear combination of their first p vectors is a minimum for a given p<w; (p=1 in figure 2). It has been shown that in most cases a good approximation of data is achieved for p<<w. Since the w PC's form a basis in R^W space, they can replicate exactly any of the n points in the set, using the scores as weights. López (1997) claims that some of the scores contain essential information on the structure of the cloud, while others are more related to noise. Following Hawkins (1974) he suggested a rule to identify the noisy scores. Once identified, such scores were used to pinpoint those points in R^W space which are prone to hold an error.

However this is not the complete answer to the problem because each point depends on w elevation values. Which one is wrong?. Once a point in R^W space is selected, the elevation (or elevations) which make it unusual should be indicated. This is done using a weighted sum of the squared scores which are related to noise. Such statistics have been suggested for the first time by Hawkins (1974). It is a semi-distance, closely related to the Mahalanobis distance. Its sensitivity in terms of the elevations values is calculated and those elevations which generate the most important contribution to the distance value are considered as errors. The calculations are carried out independently for each outlying point in the R^W space.

We have briefly presented the procedure to find an error in a single strip. The method can be applied for all row-wise strips to cover the entire DEM. The candidates obtained can be grouped and designated here in after as row-wise candidates. However, the same procedure can be applied to column-wise strips, and a different set of column-wise candidates can be obtained.

The candidates belonging to both sets (row-wise and column-wise) represent the final result. The procedure can be applied iteratively, since, once an error is detected and "corrected", the cloud is modified to some extent, and so are the scores. We keep track of the point already checked in order to avoid to select them twice; we form the candidate set as the intersection of all previous row-wise candidates and all previous column-wise candidates.

The procedure involves five actions, and it can be outlined as follows:

```
Given a DEM as a matrix of size m*n
subdivide the DEM in row-wise and column-wise strips of width w
repeat until criteria are satisfied:
    a) increment the previous  row-wise candidate set:
        a.1.- locate the columns likely to have candidates
        a.2.- within each column, find the rows that identify the candidates
    b) increment the previous column-wise candidate set:
        b.1.-locate the rows likely to have candidates
        b.2.- within each row, find the columns that identify the candidates
    c) intersect both sets
    d) evaluate criteria
    e) correct all errors
end
```

Some remarks follows. In the pseudo code we have used a single strip width w for rows and columns. This simplifies the tuning process, as will be shown later.

The process is supposed to stop when some criterion is fulfilled. For his experiment, López (1997) suggested to stop if the type I error is too big (defined as the probability of missclassify as error a good value). This criterion is useless for real errors and as will be shown below. Each iteration will be named "step" in the following discussion.

This procedure is more complex than the one of Felicísimo (1994), but it does not require that adjacent profiles appear "together". We will discuss this further in the next paragraphs.

2.3 *The modified version*

This variant has been specially designed in order to handle the problem of heavily correlated errors in space. Notice that the procedure of López (1997) has been tested with synthetic, weakly correlated errors. Its performance decays as the correlation increases. The procedure of Felicísimo suffers from the same problem, since the error at i,j is highly correlated with the one at the immediate neighbors. López's procedure does not require that the *along the strip* profiles are contiguous. Therefore we can skip some of them (the ones most correlated) for the analysis. The strip is chosen as before, but in the calculations we consider subsets created using every k-th row, k being related to the *range*, a geostatistical property (Samper *et al.* 1990) of the error field. In this paper we assume that the range can be estimated from an independent analysis: it might depend on DEM characteristics, method for obtaining it, scale of aerial photography, etc. The modified method resembles the multigrid approach (Strang 1989) used in scientific computing packages for the solution of differential equation.

3. The experiment

To test the method with real data we have selected two DEM of the Aix-en-Provence region in the South of France, both of 12.42 km by 6.9 km, 30 m spacing. A subset of 360 rows and 216 columns was used for all calculations. Both DEM have been described elsewhere (Day *et al.* 1988), and include as a significant feature Mount Sainte Victoire. The first DEM has been produced by photogrammetric measurement of spot elevations from aerial photography. Its accuracy has been estimated by multiple set-up and observation of several blocks within the DEM. An analysis of 830 duplicate points (i.e. set up and measured twice) is presented in table 1. The second DEM has been derived from a set of three SPOT images using an stereo matcher. It has been interpolated to a 30 m grid by using values within a window of size 21 pixels. Elevation values have been obtained using kriging with a spheric variogram of 4000 m^2 sill and 3000 m range, assuming an accuracy for the window of 11 m S.D. Table 2 shows the statistics for the difference between the interpolated DEM (obtained from the stereo matcher's output) and the one manually generated.

Figure 3 illustrates the main features of the DEM, and figure 4 shows the probability density function of the differences in elevation. It should be noticed that the probability of exactly zero error is negligible: only 6 out 95865 points have exactly the same elevation in both datasets.

Table 1 Comparison of 830 duplicate points of the manually derived DEM (From Day *et al.* 1988)

Mean abs. error	-0.026 m
S.D. error	1.837 m
RMSE	12.70 m
Max. (abs. size)	14.66 m
$\|error-\mu\|>3\sigma$	1.7 %

Table 2 Comparison of 95865 points of the SPOT derived DEM against the manually derived one

Mean abs. error	0.93 m
S.D. error	12.67 m
RMSE	12.70 m
Max.	193.83 m
Min.	-86.22 m
$\|error-\mu\|>3\sigma$	1.43 %

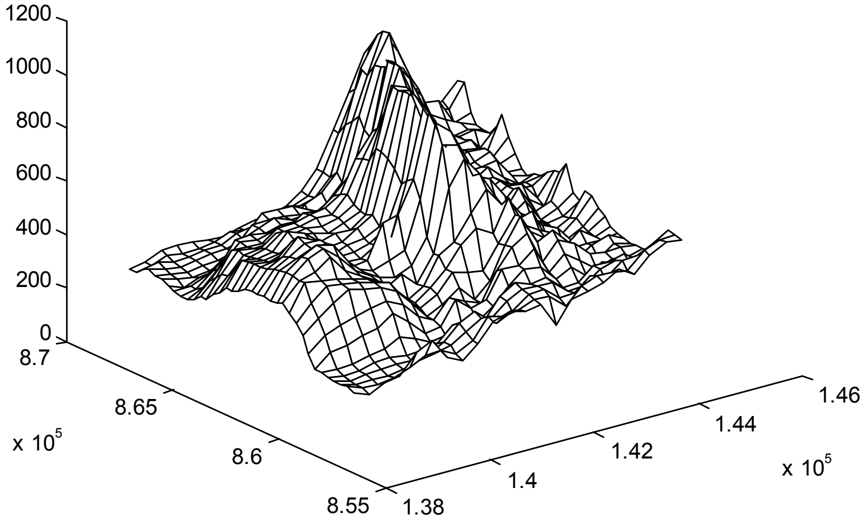

Figure 3 Illustration of the test DEM obtained using only every tenth grid value.

This leads to a paradox: since any choice for the locations will succeed in pointing a true error, the error type I will be identically zero disregarding the procedure, and the error type II will decrease linearly with the effort. This preclude to compare results with those presented in López (1997), since the author used such statistics to decide whether to stop the procedure or to continue. For real datasets a possible measure will be the RMSE between the original and the corrected elevations, and the procedure might go on as long as the RMSE exceeds a preset threshold.

Another interesting result regards some properties of the discrepancy field, i.e., the difference between both DEM. We used geostatistical techniques (Samper *et al.* 1990; Cressie 1993) to describe it. Figure 5 shows a plot of the sampled variogram. Even though the goal of the present paper is not to model the variogram itself, it can be noticed that the range can be roughly estimated as 300 m, i.e., 10 times the grid spacing. This numerical result is in agreement with results obtained by visual analysis of the discrepancy field and can be interpreted as a measure of the spatial correlation of the error field. Clearly, the occurrence of errors cannot be regarded as a local phenomenon, a hypothesis assumed by Felicísimo (1994) and López (1997).

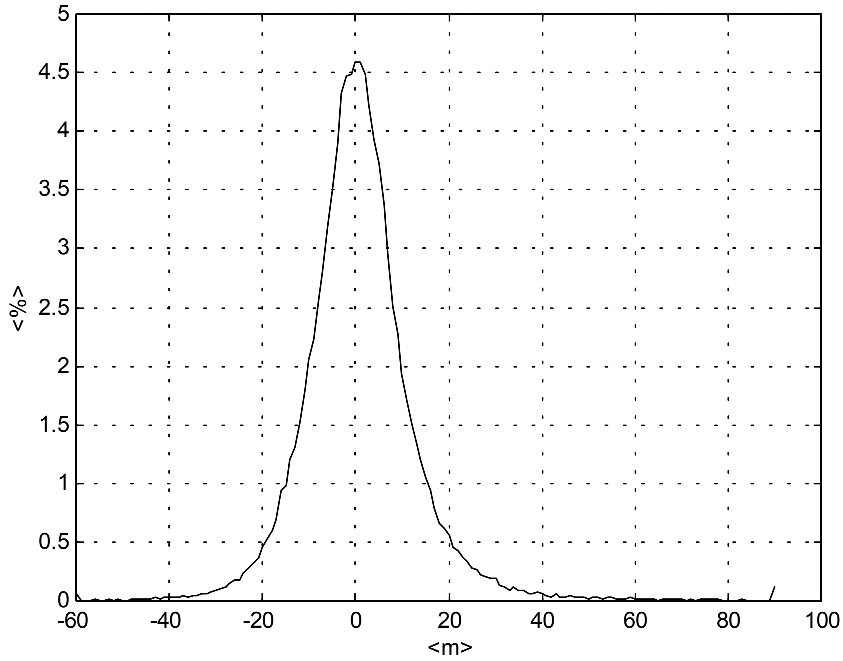

Figure 4 Sampled probability density function of the discrepancies in elevation between both DEM

Most errors are found in a smooth neighborhood regardless if they occur along breaklines or as isolated values. The errors typically influence the data over a distance of 10 pixels. At breaklines the decay should be considered across the line.

Before analyzing the accuracy results, we want to go a bit further into some hypothesis by Felicísimo (1992) a) Gaussian distribution of the errors and b) relationship between outlying values of the $t_{i,j}$ population and the true errors. We show in figure 6 a QQ-plot of the distribution of the original $t_{i,j}$ population. The QQ-plot produces a linear relationship when two distributions are of the same type (even though with different parameters). In this case the target distribution is Gaussian. The x-coordinate corresponds to the normal cumulative error function, with no units, while the y-coordinate is the sampled cumulative density function of $\delta_{i,j}$, measured in m. As it can be seen, Gaussian distribution is hardly achieved.

On the other hand figure 7 shows a QQ-plot comparing the distribution of the $\delta_{i,j}$ against the real errors. In this case the similarities are evident, and this might lead to the wrong conclusion that $\delta_{i,j}$ is heavily correlated with real errors (which in turn offers a tool to locate big errors finding big values of $\delta_{i,j}$). Unfortunately this is not the case. What can be concluded from figure 7 is that both populations belong to the same (unknown) class, but they might be completely independent. In fact the linear correlation coefficient is 0.0858.

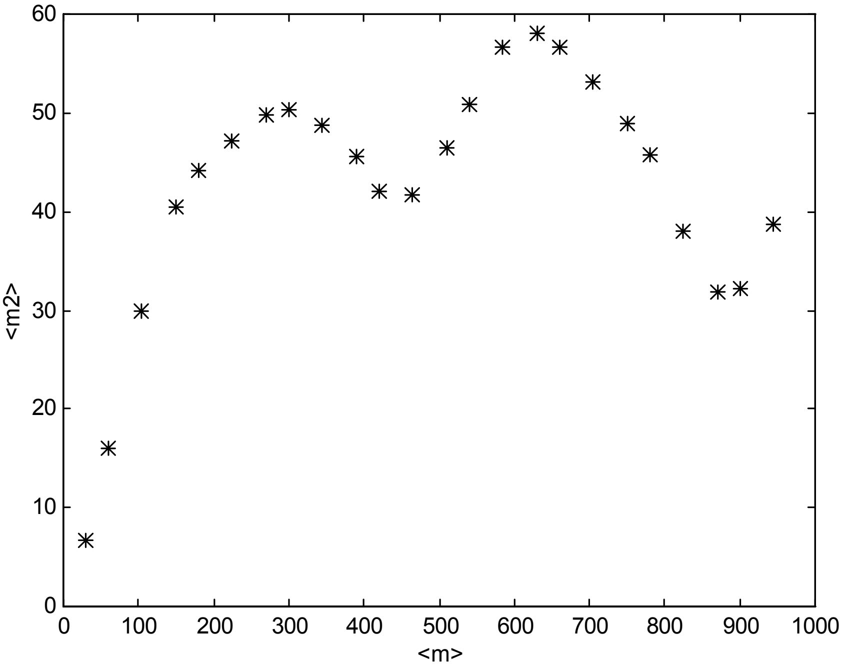

Figure 5 Sampled variogram of the difference of both DEM vs. distance. Results obtained from the GEOEAS software

4. Results

We may analyze the results using different statistics. The most interesting one will take into account the evolution of the elevation accuracy in terms of the editing effort. For our purposes we measure the editing effort as the number of elevation values checked divided by the total number of points in the DEM. We assumed that the user has a correction procedure and that procedure is perfect.

Normally the accuracy of a DEM is not directly known to the user; it can be estimated through sampling in isolated points if more precise measurements are available.

For practical purposes it might be more meaningful to use statistics from the distribution of the errors detected while working with the dataset For example, its RMSE will measure the size of the errors detected by the method for a given effort. We disregard the RMSE for each step, because its variability precludes any simple analysis.

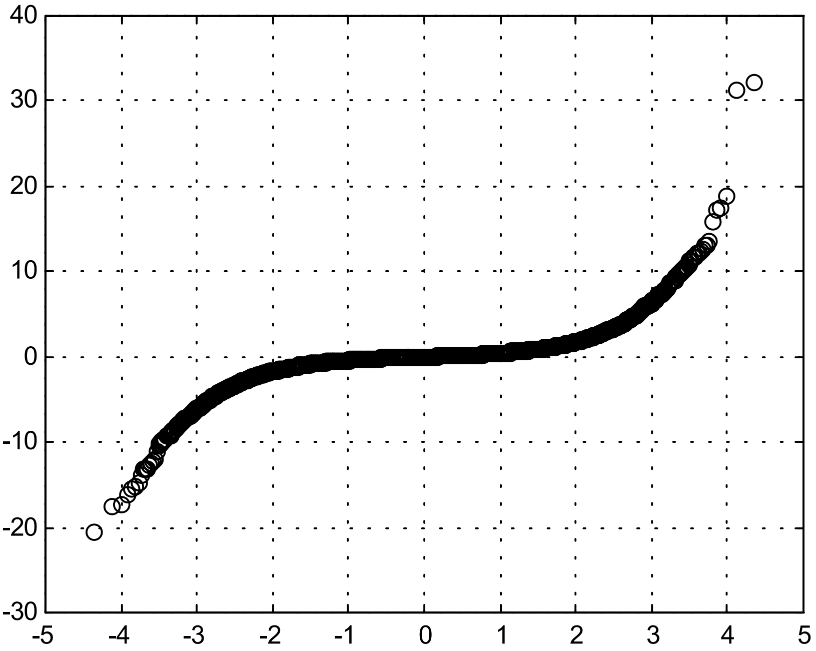

Figure 6 QQ-plot of the N(0,1) cumulative density function vs. the sampled cumulative density function of $\delta_{i,j}$, for the available DEM.

We believe that a clear measure of the effort involved should be included. The effort per step (in turn) depends strongly on the choice of the margin level. It regulates how much of the tail of the distribution of the noisy scores will be regarded as being in error. Cutting out the tails might produce an empty set of candidates. In order to avoid this we slightly increased the margin level to assure that there will be candidates to check in each step.

Figure 8 shows the evolution of the accuracy measured in terms of the RMSE for a strip width w=8. The boundaries of the dashed regions at the top and the bottom show the *worst* and *best* possible operation locus. The former is obtained by considering first the smallest discrepancies, while the latter corresponds to selecting the largest discrepancies first. Under our assumptions both lines should meet at 0 and at 100 per cent. Even though both limits are hardly of practical interest (because it requires knowing the errors in advance) they give a better understanding of the process. Lines with the -o- symbol are for the Felicísimo (1994) method while the others are for different controlled scores as of López (1997) method. Figure 8a has more detail in the low effort region, while figure 8b has been extended up to 15 per cent effort. It is clear that the Felicísimo's method outperforms the López's method in the long run, but at the low effort they are similar. This region is of primary importance for two reasons. Firstly because most users will not go too much further. End users neither have extra data nor too much tools, so they will correct at most the worst errors. DEM producers might go back and make another measurement, but this might become a boring task if new values do not differ substantially from the old ones. Secondly, according to Torlegård *et al.* (1986) blunders typically account for less than 3 per cent of the dataset, 0.5 per cent being a median value. Thus pursuing the task over such limit might be misleading, because the methods have been designed for finding gross errors only.

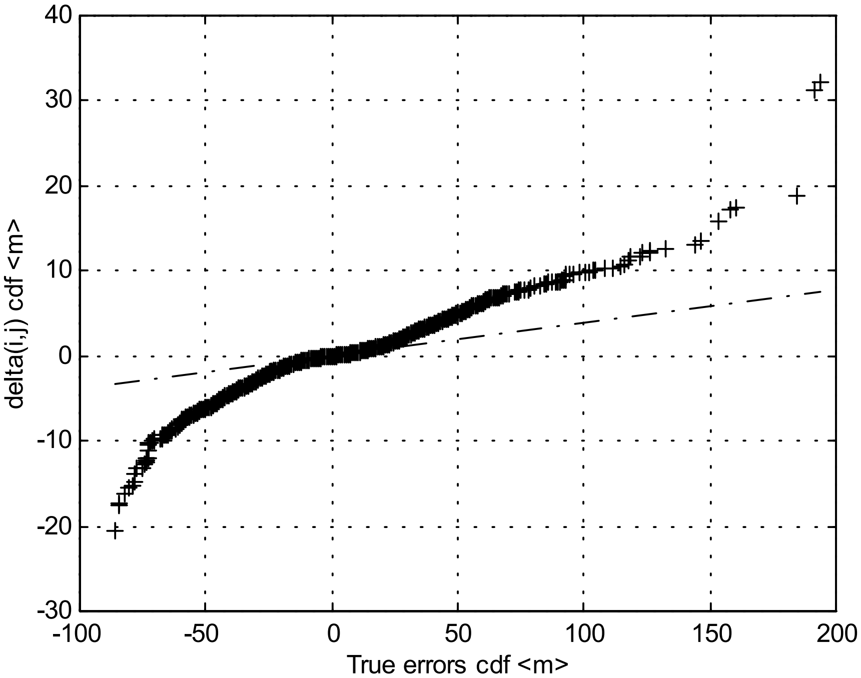

Figure 7 QQ-plot of the sampled cumulative density function (cdf) of the true errors vs. the $\delta_{i,j}$ cdf for the available DEM.

It should be noticed that none of the methods shows at 0 per cent effort a slope comparable to the best possible method, which implies that the most important errors are not found in the early stages of the procedure.

We also tested some other options for the width parameter w which will not be presented here. Figure 9 compares the accuracy performance of the modified method vs. Felicísimo's one. The figure was obtained after subdividing the DEM in regions of width 72 rows, and building the strips taking every 9th row within the region. Thus the "strip" width is again 8. Notice that we skip nearly 10 rows, as suggested by the range of the variogram. The plot of the Felicísimo's method is again included for comparison. The most striking fact is the difference in the slope at 0 per cent effort which is markedly closer to the best one. This implies that larger errors are found earlier, leading to a faster decrease of the RMSE. However, once those important errors are removed, the remaining ones are difficult to locate, and the simpler Felicísimo's method is better if the effort exceeds 1.75 per cent.

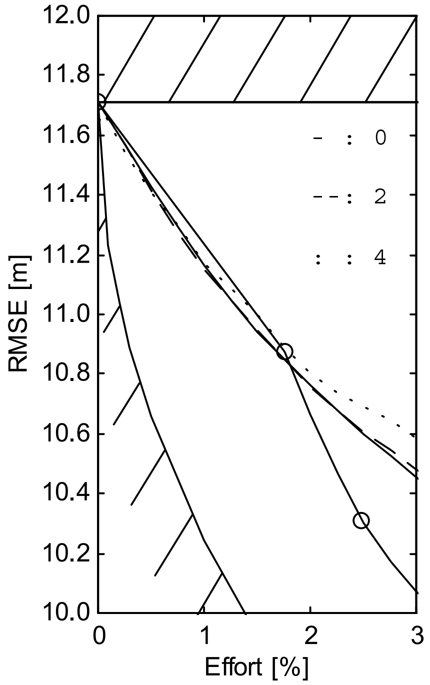

Figure 8 Evolution of the accuracy (measured by the RMSE in m) vs. the effort for the methods of Felicísimo (1994) (with the -o- symbol) and López (1997). Results for w=8. Different lines correspond to different number of uncontrolled scores. Left plot 8a shows details of the right one 8b.

The end user can calculate RMS values of the errors already found like those presented in figure 10. The x-coordinate is the effort defined as before, while the y-coordinate is the RMSE of the population already corrected. The 0 per cent value is not defined. Plots correspond to the Felicísimo (1994) approach and the modified method of López (1997). It is clear that the former finds larger errors in the "long" run (over 1.75 per cent effort) but the latter is fairly better for lower effort values. Three lines with different number of uncontrolled scores are shown, and it is clear that the one of 0 value is very similar to the one of 2, except very close to the 0 per cent effort.

We also analyzed the spatial location of the errors found when a substantial amount of work has been done. Figure 11 shows the places where Felicísimo's method pointed out the errors up to the 3 per cent effort (in black), and up to 15 per cent (in gray). We noticed that most of them are concentrated along significant features of the DEM, namely breaklines where slope changes abruptly. In such points the second order polynomial is not a good approximation of the surface, so differences larger than expected arise. Once some values are corrected, such differences are even more evident, but since we do not allow any point to be corrected twice, its closer neighbors become candidates, explaining the "clear" image. Figure 12 shows the pattern for the modified method of López. The image looks "noisy" since points are located randomly. Due to space limitations we cannot go further in the comparison of both patterns.

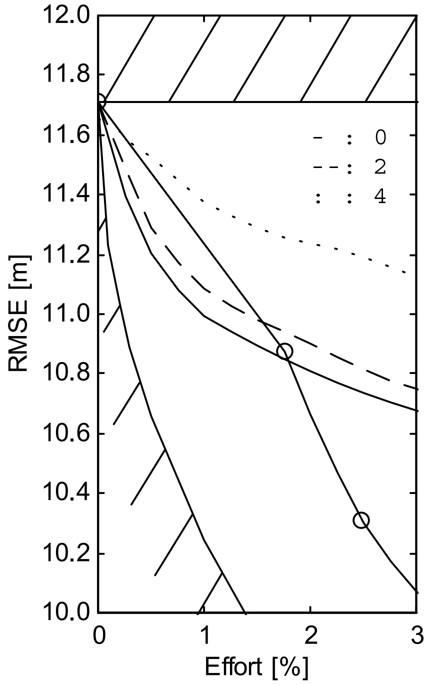

Figure 9 Evolution of the accuracy (measured by the RMSE in m) vs. the effort for the methods of Felicísimo (1994) (with the -o- symbol) and the modified of López (1997). Results choosing every 9th row. Different lines correspond to different number of uncontrolled scores. Left plot 9a shows details of the right one 9b.

5. Discussion

We have compared two published methods for locating errors (also named outliers or blunders) in raster datasets. We also suggested a modification for one of them, and we carried out a comparative test for all three methods using real data with known errors. The method suggested by Felicísimo (1994) is very simple, but no results using either synthetic or real errors were previously reported. One interesting fact is that this method is parameter free. However it has been derived under some hypotheses that do not apply to the DEM used in this study. It relies on a low order polynomial interpolator using only nearest neighbors. We think that it will work better in smooth terrain. The use of low order polynomials tends to pinpoint locations which are close to each other, a situation which is more likely to occur with systematic errors. For further work we suggest considering the use of a local Universal Kriging interpolator (Samper *et al.* 1990) using more neighbors, which is in line with the findings of Giles *et al.* (1996) who also used a window with 11 by 11 elements. The Kriging approach also allows to model different spatial correlation scales.

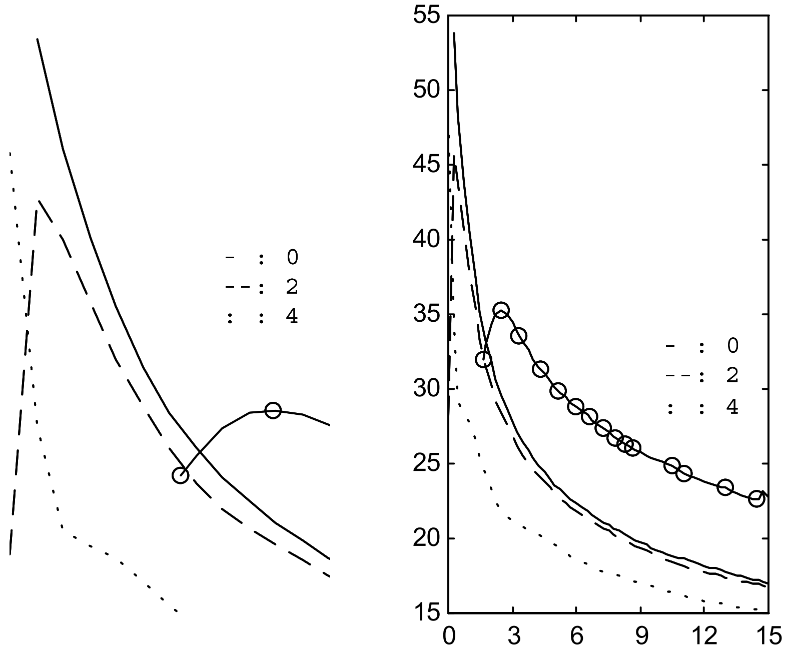

Figure 10 Evolution of the RMSE found of the cumulated errors up to a given effort vs. the effort, for the methods of Felicísimo (1994) (with the -o- symbol) and the modified of López (1997). Results choosing every 9th row, resulting strips of w=8. Left plot shows details of the right one

The overall results show that *if* better elevation values can be derived using the same raw data, this approach leads to higher accuracy, provided that they there are no systematic errors.

The method outlined by López (1997) has been designed for and tested with synthetic errors with very low spatial correlation. For our case, where errors show heavy spatial correlation, it performs only slightly better than Felicísimo's for low effort, but it is outperformed in any other case. We consider the performance of López's method as poor.

In order to handle the spatial correlation of errors, we have proposed a modification of the method by López. We form the strips by subsampling the DEM at each k-th row. From a programming point of view this is a minor change. In real applications, the number k has to be fixed *a priori*. Östman (1987b) suggested that k is strongly connected with both the DEM and the acquisition method. We estimated the range from the sampled variogram. López (1997) describes a rule how to determine how many scores are considered describing the structure of the cloud. This rule suggest a value of 2. Slightly better results were obtained using 0. However, in a first approximation the rule is still valid.

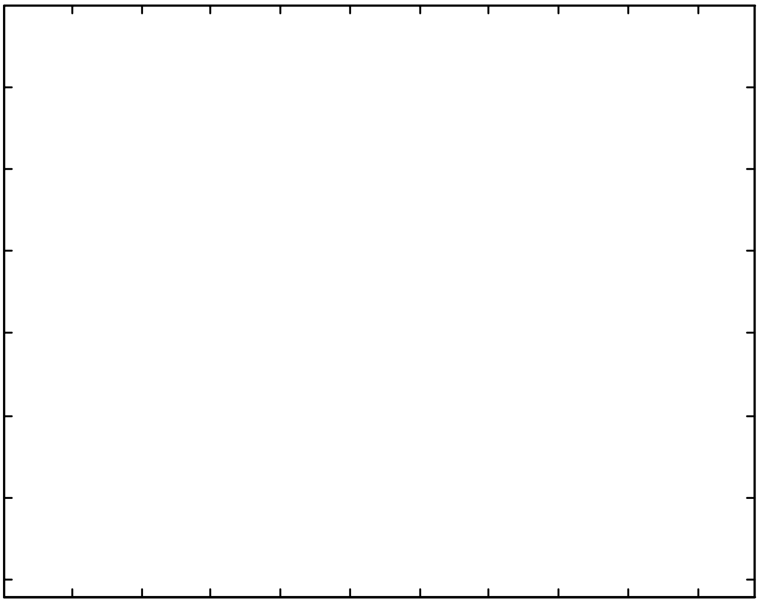

Figure 11 Binary map of the errors located up to the 15 per cent effort with the method of Felicísimo (1994). Black areas are for the suggested locations up to the 3 per cent effort; gray ones are obtained after 15 per cent effort

All three methods have been used in an iterative fashion. Once some errors were removed, all the calculations have been carried out again, and new candidates appear. If this is not the case, some parameters are modified automatically (lowering confidence limits, for example) in order to continue the operation. We continue until 15 per cent of the DEM elevation values have been corrected or confirmed. According to Torlegård *et al.* (1986) gross errors account for less than 3 per cent of the population, so the 15 per cent limit is well within either the systematic (as defined by Thapa *et al.* 1992) or the random error set, provided the first 3 per cent were really gross errors.

We assumed that, once an error is located, it can be replaced by a "better" value. In real applications the procedure will be different depending on the user. In a DEM production environment, some action can be taken to check these identified isolated values. In photogrammetric measurements these checks can be done before removing the stereopair. The goal here is to improve the overall accuracy, while the effort is less crucial. On the other hand, the end user is left alone in most cases, because he may not be able to go to the original data sources. Therefore he will be interested in "evident" errors, i.e. those of relevant size (which are typically few).

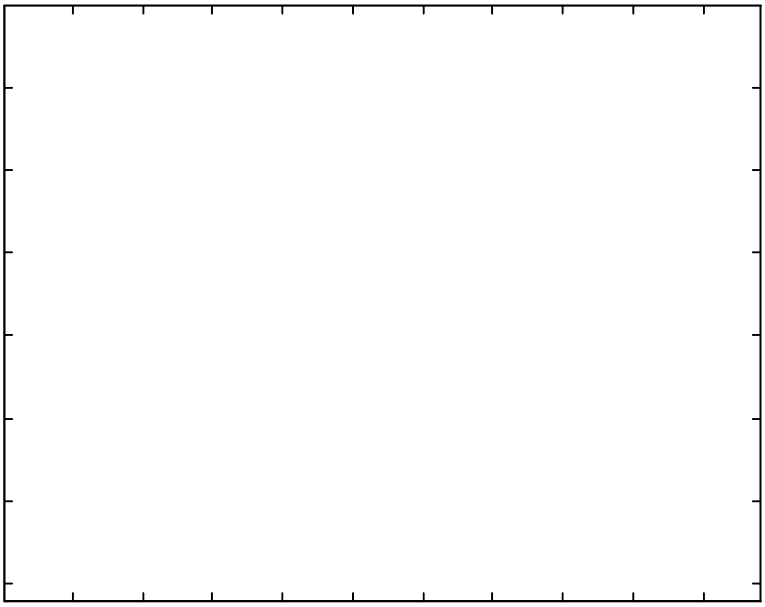

Figure 12 Binary map of the errors located up to the 15 per cent effort with the modified method of López (1997). Black areas are for the suggested locations up to the 3 per cent effort; gray ones are obtained after 15 per cent effort

A comment about the computer time requirements: the method of Felicísimo (1994) is fairly cheap (of the order of m.n operations, being (m,n) the size of the DEM), while the method of López (1997) and the modified procedure presented here involve, for each step, the computation of (m/w).(n/w) covariance matrices of size (w,w), which takes $[(n/w).O(n^2)+(m/w).O(m^2)].O(w^2)$ operations; to calculate the eigenvectors requires in turn $[(n/w)+(m/w)].O(w^2)$ operations, and to project each strip to calculate the scores requires (m+n).w operations. Some other operations are required but depend linearly on m and n. In our example, a DEM of size m=360 and n=216, for w=8, about 5 minutes per step are required using MATLAB in a SUN Sparc 20. The overall procedure is considered cheap in terms of computer time.

6. Conclusions

Some methods to locate gross errors in quantitative raster data have been presented, and they were tested using a grid-based DEM with known errors. The DEM, derived from SPOT data, has elevations ranging from 181 to 1044 m. A more accurate DEM of the same area is available; it has been considered as the ground truth. The hypothesis of errors uncorrelated in space seems to be wrong at least for this case, as well as the assumption of gaussian distribution for the residuals. This poses serious concern about the usefulness of some previously published algorithms (Felicísimo 1994; López 1997) and motivated this work.

The results suggest that Felicísimo (1994) method find mostly what is regarded as systematic errors, mainly due to the interpolation algorithm (biquadratic polynomial). López (1997) show similar results in terms of RMS of errors only in the early stages of the correction process.

In order to handle the significant spatial correlation observed a modified version of the method of López (1997) has been designed and tested with the same dataset. The results were significantly improved and exceeded those of Felicísimo up to a certain level of effort, the effort being defined as the fraction of the DEM elevations corrected or revised. This effort level (1.75 per cent) is of the order of the number of gross errors typically found in DEM; moreover its location pattern looks sparse and random, as opposed to the pattern produced by Felicísimo's method.

The modified method has some free parameters; the most important is an estimate of the correlation lag (or the range of the variogram). It can be estimated from a limited number of independent control points; some authors claim that such value depends on the method for acquisition of the DEM and the DEM itself.

We assumed that once an error is identified, it can be corrected. In the case of using the algorithm in a semi-automatic production environment, the method warns the operator about possible errors before the stereopair is unmounted, enabling new measurements. In a fully digital production environment, some correlation thresholds are usually varied to minimize computer time. The method may be used to selectively strengthen the correlation thresholds in suspicious points. In case there is no possibility to verify the errors, e.g. for end users, the algorithm will help to locate the most unlikely values; they may be replaced with the aid of some suitable interpolation method. If there are some independent sources (maps, etc.) they could be used for checking.

7. Acknowledgments

This work was started at the Department of Geodesy and Photogrammetry, Division of Geoinformatics, of the Royal Institute of Technology, Stockholm, Sweden, under the supervision of Assoc. Prof. Hans Hauska. Prof. Jan Peter Muller from the University College, London, have kindly supplied both DEM and related information. The author wishes to acknowledge BITS (Swedish Agency for International Technical and Economical Cooperation) and CONICYT (Consejo Nacional de Investigaciones Científicas y Técnicas, Uruguay) for partial funding. Computer resources have also been kindly supplied by the Centro de Cálculo of the School of Engineering, University of the Republic, Montevideo, Uruguay.

References

ABREU, E.; LIGHSTONE, M.; MITRA, S.K. AND ARAKAWA, K., 1996, A new efficient approach for the removal of impulse noise from highly corrupted images. *IEEE transactions on image processing*, **5**, 6, 1012-1025.

ACKERMANN, F., 1995, Digitale Photogrammetrie - Ein Paradigma-Sprung. *Zeitschrift fur Photogrammetrie und Fernerkundung*, 3/95, 106-115

ANONYMOUS, 1994, ESRI tops the charts again. *ESRI ARC News*, **16,** 3, 35-35

BETHEL, J. S.; MIKHAIL, E. M., 1984, Terrain surface approximation and on-line quality assessment. In *Proceedings of the 15th. International Congress of the International Society for Photogrammetry and Remote Sensing, Rio de Janeiro. Commission III*, **25**(A3a), 23-32

CRESSIE, N., 1993, Statistics for spatial data. John Wiley and Sons, ISBN 0-471-00255-0, 900 pp

DAY, T.; MULLER, J.P., 1988, Quality assessment of Digital Elevation Models produced by automatic stereo matchers from SPOT image pairs. In *Proceedings of the 16th. International Congress of the International Society for Photogrammetry and Remote Sensing, Kyoto*. Commission III, 148-159.

FELICÍSIMO, A., 1994, Parametric statistical method for error detection in digital elevation models. *ISPRS J. of Photogrammetry and Remote Sensing*, **49**(4): 29-33

FÖRSTNER, W., 1983, On the morphological Quality of digital elevation models. In *International colloquium on Mathematical aspects of Digital Elevation Models, Stockholm*.

GILES, P.T. AND FRANKLIN, S.E., 1996, Comparison of derivative topographic surfaces of a DEM generated from stereoscopic SPOT images with field measurements *PE&RS*. **62**, 10, 1165-1171

HADI, A.. S., 1992, Identifying Multiple Outliers in Multivariate Data. *J. Royal Statist. Soc. B* **54**, 3, 761-771

HADI, A. S., 1994, A Modification of a Method for the detection of Outliers in Multivariate Samples. *J. Royal Statist. Soc. B* **56**, 2, 393-396

HANNAH, M. J., 1981, Error detection and correction in Digital Terrain Models. *PE&RS* **47**, 1, 63-69

HAWKINS D. M., 1974, The detection of errors in multivariate data, using Principal Components. *JASA*, **69**, 340-344.

HAWKINS, D. M., 1993a, A feasible solution algorithm for the Minimum Volume Ellipsoid Estimator in Multivariate data. *Computational Statistics* **8**, 95-107

HAWKINS, D. M., 1993b, The feasible set algorithm for least median of squares regression. *Computational Statistics & Data analysis* **16**, 81-101

HUNTER, G. AND GOODCHILD, M., 1996, Comunicating uncertainty in spatial databases. *Transactions in GIS* **1,** 1, 13-24

JOHN, S. A., 1993, Data integration in a GIS - The question of data quality. *ASLIB Proceedings*, **45**, 4, 109-119.

LEBART, L.; MORINEAU, A.; TABARD, N., 1977, Techniques de la description statistique: Methodes et logiciels pour l'analyse des grands tableaux. Ed. Dunod, Paris, 344 pp.

LI, Z., 1992, Variation of the accuracy of digital terrain models with sampling interval. *Photogrammetric Record*, **14**(79), 113-128

LÓPEZ, C., 1997, Locating some types of random errors. To appear in *IJGIS*

ÖSTMAN, A., 1987a, Quality control of Photogrammetrically sampled Digital Elevation Models. *Photogrammetric Record*, **12**(69), 333-341.

ÖSTMAN, A., 1987b, Accuracy estimation of Digital Elevation Data Banks. *Photogrammetric Engineering and Remote Sensing*, **53,** 4, 425-430.

MITRA, S.K. AND YU, T.H., 1994, A new nonlinear algorithm for the removal of impulse noise from highly corrupted images. *IEEE 1994 International symposium on circuits & systems* **3**, 17-20

NEBERT, D., 1995, Status of the National Geospatial Data Clearinghouse on the Internet. In *Proc. of the 15th. Annual ESRI User Conf., May 1995*. Also http://www.fgdc.gov/clearinghouse/pubs/esri95/p196.html

NEBERT, D., 1996, Supporting Search for Spatial Data on the Internet: What it means to be a Clearinghouse node. *In Proc. of the 16th Annual ESRI User Conf.* Also http://www.fgdc.gov/clearinghouse/pubs/esri96/revised.html

RICHMAN, M. B., 1986, Review article: Rotation of principal components. *J. of Climatology*, **6**, 293-335.

SAMPER, F.J. AND CARRERA, J., 1990, Geoestadística: aplicaciones a la hidrología subterránea. ISBN 84-404-6045-7 480 pp (in spanish)

STRANG, G., 1989, Introduction to applied mathematics. Wellesley-Cambridge Press, 510 pp

THAPA, K. AND BOSSLER, J., 1992, Review article: Accuracy of spatial data used in Geographic Information Systems. *Photogrammetric Engineering & Remote Sensing*, **58**, 6, 835-841

TORLEGÅRD, K; ÖSTMAN, A. AND LINDGREN, R., 1986, A comparative test of photogrammetrically sampled digital elevation models. *Photogrammetria*, **41**(1), 1-16

Appendix 6

López C., 1996, "Improvements over the duplicate performance method for outlier detection in categorical multivariate surveys" *Journal of the Italian Statistical Society, **5**, 2, 211-228*

IMPROVEMENTS OVER THE DUPLICATE PERFORMANCE METHOD FOR OUTLIER DETECTION IN CATEGORICAL MULTIVARIATE SURVEYS[1]

CARLOS LÓPEZ
Ingenieros Consultores Asociados
Cerro Largo 1321, Montevideo, Uruguay

Abstract: The detection of errors and outliers is an important step in data processing, specially those errors arising from the data entry operations because they are of the entire responsability of the data processing staff. The duplicate performance method is commonly used as an attempt to detect such type of errors. It implies typically typing twice the same data without any special precedence. If the errors are uniformly distributed among individuals, retyping a fraction of the total will also remove typically the same fraction of the errors. A new method which is able to improve that procedure by sorting the records putting first the most unlikely ones is presented. The ability of the present methodology has been tested by a Monte Carlo simulation, using an existing database of categorical answers of housing characteristics in Uruguay. At first, it has been randomly contaminated, and after that, the proposed procedure applied. The results show that if a partial retyping is done following the proposed order about 50% of the errors can be removed while keeping the retyping effort between 4 and 14% of the dataset, while to attain a similar result with the standard methodology 50% (on average) of the database should be processed. The new ordering is based upon the unrotated Principal Component Analysis (PCA) transformation of the previously coded data. No special shape of the multivariate distribution function is assumed or required.

Some keywords: Data checking, Census data management, Outlier detection, Principal Component analysis, Categorical data

I.- Introduction

A recurring problem in the creation or maintenance of a large computerized data base is the correctness of the information entering the base. If high volumes of data are involved, then data entry operation tends to be carried out by less qualified personnel, and verification is less extensive. Thus, action is required to maintain the base's integrity, and the fact that large volumes of machine-readable material are involved suggests that, as far as possible, this screening action should be automated. Clearly typing errors is not the single source of errors existing at the machine level; however on principle they can be kept under control.
There are many classical examples of typing errors, even from the early days of computer development. A classic one is described by Coale *et al.* (1962) who reported an error in the 1950 U.S. census figures that resulted when a small fraction of computer cards were

[1] Published in the Journal of the Italian Statistical Society, 1996, **5**, 2, 211-228

punched one column to the right of the proper position, so that an unusually high number of 14-year-old widows was reported. Only after discovering the reason for that error, they were able to anticipate errors in the age-distribution of Indians individuals. Even though the total amount of erroneous records was fairly small (below 1/100 of one per cent) certainly rare categories might be greatly affected by such spurious cases. Notice that all the fields in the record have values within their own feasible range.

A general procedure for locating typing errors in a data entry process is the duplicate performance method. If a data typing operation is performed twice, independently, and if the results are compared by a method that can be assumed error free (such as a computer program comparing files after data entry), and if all the disagreements are corrected, then the only errors remaining in the data set are those where both staff members were in error. If the ratio of disagreements to total items is low, then the individual error rates of both persons are low, and the probability of joint errors (the product of the probabilities of individual errors) is lower still (Strayhorn (1990)). The method is extremely simple, and it applies for any kind of data, both quantitative or categorical. Despite its simplicity, it has some desired properties: the probability of locating an error is independent of the error itself, so trivial errors will be corrected as well as subtle ones. This will help in keeping the statistical properties of the database. It is also independent of the *order* the retyping is performed, so in principle, if only a fraction of the dataset is retyped, typically the same fraction of the errors will be corrected. Another advantage is that the procedure does not require a large database, so it can be applied also to small ones.

The literature about editing survey data is considerable, but somewhat scarce regarding quality control of categorical data. Fellegi *et al.* (1976) presented a methodology specifically suitable for qualitative or categorical data. It is based upon the existence of rules which relate the different fields in each record. Such rules should be given by experts, and express the judgment of them that certain combinations of values or code values in different fields are unacceptable. If a particular record does not satisfy one or more of those rules, the field (or fields) that contribute to them are rejected or modified in order to attain a feasible record. Notice that this procedure relies on the existence of explicit rules (and experts behind them) and requires some manipulation of the rules before application. No experimental results are presented in the paper.

Paradice *et. al*. (1991) presented a methodology for controlling *incoming* data to a database. Their approach focuses in minimizing the time a wrong record stays in the system, basically by limiting its chances to pass some logical tests created by experts, and tailored for the particular application. Not all the attributes of a record are important for all applications, so new tests may be required for different users of the same data. For the applications the authors are involved in, individual records should be handled also individually and not "in aggregate" so errors will have significant effect for one particular record, but possible not for the whole database. The paper also gives a performance evaluator for the overall error diagnostic procedure, which gives an enterprise measure of success. They claim this benchmark gives a clear measure for evaluating current verification procedures and proposed changes. Even though we could not apply this methodology for an already existent database, or even one that is created in a single task (a national census, for example) it will give us the chance to qualify the procedures used in a continuously updated process (like economical data).

Apart from the methods specially devised for categorical data, we want to mention some of the methods available for quantitative data, since we will adapt some of them for the former case. Typically the authors rely on assumptions about the data distribution. For example, Little *et. al*. (1987) presented a methodology based upon multivariate normality of the data. They used a log-transformed population, and look for linear relationships between the new variables. Using the squared Mahalanobis distance as an estimator, the author analyzed its sampled distribution exploring graphically the departure from a transformed chi-squared distribution. All instances that renders values that are "far" in some sense to the theoretically assumed behavior are flagged and edited by experts. They also extend their methodology for incomplete datasets, limiting for each individual the Malahanobis distance to the available data.

A related approach has been presented by Hawkins (1974) based upon Principal Component Analysis (PCA) of the data. Instead of using the Mahalanobis distance, he proposed to use other statistics which are intended to be more sensitive and to have better performance when compared with standard statistics (χ^2, etc). However, some problems arise while calculating the eigenvalues of the covariance matrix in real data. The existence of outliers may affect its values, so more robust procedures should be preferred, and not all the data can be regarded as normally distributed.

López *et al*. (1994) presented a methodology that overcomes some of these drawbacks. Instead of using the distribution of a single number like the Mahalanobis distance or the Hawkins´s statistic for flagging an instance, they proposes to use k independent tests applied over the projections of the given data on the eigenvector´s basis. No fitting with any distribution is required. The rest of the paper is devoted to show a connection of quantitative data procedures to categorical ones, and to present some simulated results.

The work is organized in nine sections. Section **I Introduction**, has discussed some work representing the state-of-the-art on the subject. Section **II Motivation and assumptions** introduces the main ideas. Section **III Experimental test design** describes the simulation carried out to examine the performance of the method with a particular dataset. Section **IV Methodology** describes the steps required to apply the procedure. Section **V Results** summarizes the success by means of some performance indicators and finally section **VI Discussion** compares the results and analyzes advantages and drawbacks, while section **VII** states the **Conclusions**. **Acknowledgments** and **References** are included as sections **VIII** and **IX**.

II.- Motivation and assumptions

For the sake of simplicity, we will assume hereinafter that by typing twice a record all errors are removed. This will help us in simplifying some arguments, and the reader will easily notice that this not a key hypothesis.

We mentioned before that the duplicate performance method ability is independent of the order the records are retyped. If we assume that the wrong individuals are uniformly distributed in the population, retyping a fraction of the dataset will most likely correct the same fraction of errors. This paper is devoted to find a reordering in the data, designed to put first the records that are prone to hold some errors, so partial retyping will eliminate more errors than without any reordering.

To do so, we will try to locate outliers in the dataset. What is meant by outlier in categorical data may differ from the concept for real-valued data. It is also assumed here that the dataset has passed successfully some trivial logical tests, which pointed out for example, more than one mark in mutually exclusive answers, or similar things. Also all the coded values are within their prescribed ranges. These logical tests are very crude, and certainly should not be confused with the edits designed by experts in the particular data (Fellegi *et. al*. (1976)). It should be regarded more as a computer specification for the data, rather than a quality control procedure.

So we will consider only the problem of selecting a specific record (a single survey in the example) on the basis that there is something in the answers that make them unusual. Such record should be retyped. Notice that this procedure will diminish the variability in the data, because “feasible” errors are prone to be ignored.

In a real processing environment, if the record is still unusual it will be carefully analyzed by a trained specialist, which may found (or not) reasons to reject or modify some answers in the particular record. This fact will not be considered here, but the methodology is in fact devoted to give the specialist a smaller selected set, with higher probability of holding true errors.

It should be stressed that errors arising from the the typing stage is one among others sources of errors; however they are important in the sense that they can be kept under control. Significant errors can be introduced in earlier stages (like the coding of non-categorical answers) which cannot be controlled by the duplicate performance method, but can be handled by the procedure to be presented below.

In categorical data, the codification procedure usually generates for each question a set of feasible values. For technical reasons, those values are frequently coded as integers, but the integer value itself is meaningless. In order to manage categorical data with PCA, one should translate such integers in a way that the results do not depend upon :

a) changing the order of the alternatives in the question
b) changing the integer codes

It will also be assumed that all the answers have the same relative importance. The technique to be presented, was designed to be applied for processing the 1996 National Census of Population and Housing in Uruguay (population ~3 million, houses ~200.000) to check only categorical answers. The data was not be typed, but scanned and processed via automatic recognition routines, handling handwritten text, number and marks. Even though automatic recognition of marks are known to be very reliable, it is intended to flag and check dubious data while keeping the manual typing effort low.

III.- Experimental test design

A Monte Carlo simulation is performed, modifying the answers of a subset of the raw data collected and processed during the 1985 National Census in Uruguay, and testing the ability of the methodology to locate them. The subset chosen reports housing characteristics in the Flores region and has been typed twice. Only private houses cases, without missing values were considered. The final set has 4963 events, but to diminish computer time requirements, the simulations were carried out over only 2500 individuals.

The dataset is claimed to be typed twice, and the original records are not available. This fact makes it difficult to properly model a pattern of "rule" for real errors, so only reasonable assumptions could be made.

In order to obtain a contaminated set, a prescribed number of records were chosen at first and then a random number generator choose a fixed number of questions (out of 20) to modify. For each of them, the existing answer was changed to a different value, but still belonging to its feasible set (assuring that they were different with the original one). That was considered a suitable choice for modeling "true" errors. The *total number of contaminated records* were fixed as 10, 5, 3 and 1.5% of the subset of 2500 individuals. The figures to be presented correspond to the 3% case, which implies 75 wrong cases.

IV.- Methodology

In this section, all steps required for processing a categorical dataset are described. Given the data, the corresponding question list and the feasible options, the user should eliminate those fields which are *a priori* uncorrelated with the others. Typical examples for survey data are all the information related with the zip code, city code, address, etc. Also numerical quantitative data should not be considered (for example: age, size of the building, etc.) except if a categorization is applied.

The dataset is usually available in table format, one individual per row, and one question per column. In order to have a numerically useful representation, we will binarize the dataset, creating a new table containing only 1 or 0. This also make the data homogeneous (dimensionless). In order to binarize the dataset, one may think on a multiple choice sheet. For any particular question, there are room to choose between some (maybe mutually exclusive) alternatives. Instead of coding a single number for the answer, we may equally store all the alternatives, putting a 1 or 0 if the option is true or not. In other terms, each column of the original table expands to as many columns as alternatives in the question, allowing only 0 or 1 as an answer. After repeated for all questions, the data are transformed into binary format, and the covariance matrix can be calculated.

Since the methodology to be applied relies upon exploiting the empirical relationships between the answers, all the questions that are weakly correlated with the other data will not be considered by this procedure. In early stages of the work it has been found that also "almost trivial" answers were a source of problems, because they behave like uncorrelated answers. For example, in the test dataset more than 96% of the population has direct connection with the electrical power supply. So the corresponding answer has trivially nothing to do with the others answers. That was also the case for the questions "do you have a freezer?", "do you have telephone at home" and others which almost always have been answered "no" in this particular dataset. So, if more than 95% of the population answers are the same for any binary option, the option will be removed for the final test. A second criteria was applied trying to eliminate uncorrelated answers. If the off-diagonal elements of the correlation matrix are very close to 0, the corresponding option is also removed. The threshold has been chosen as 10 times the machine ε. (defined as the largest number which satisfies $1+\varepsilon=1$ in finite precision arithmetic). Those questions were removed before applying the outlier detection process. The final dataset has 20 questions, with 69 alternatives (options).

To highlight unusual records, a PCA derived method is being proposed. PCA is a well known methodology that transform the original (mutually correlated) data in another uncorrelated but equivalent presentation. Usually such transformation is performed in order to reduce the dimensionality of the problem. Only the first Principal Components (PC) are retained, and most of the variance in the original set is explained through them. The remaining PC are usually neglected.

Hawkins (1974) pointed out that those neglected PC may serve as outlier detectors. PCA transforms the covariance matrix Σ to diagonal form, so $E\Sigma E^T = \Lambda$. Any instance of the data X_i is also transformed to $W_i = E(X_i - \mu)$, being μ its sampled mean value. Obviously the elements w_{ij} are a linear combination of the components of X_i. The $w_j(X_i) = w_{ij}$ components are mutually uncorrelated, and have variance λ_j (the associated eigenvalue). The PCA residual test statistic is defined by Hawkins as

$$T_2(X_i) = \sum_{j=1}^{k} \frac{1}{\lambda_j} w_{ij}^2$$

being k limited for only *some* of the terms in the vector w_i. When the summation takes place over *all* the terms, this statistic equals the Mahalanobis distance, defined as

$$\Delta_i^2 = (X_i - \mu)^T \Sigma^{-1} (X_i - \mu) = (X_i - \mu)^T \left(E^T \Lambda E\right)^{-1} (X_i - \mu) =$$

$$(X_i - \mu)^T E^T \Lambda^{-1} E(X_i - \mu) = W_i^T \Lambda^{-1} W_i = \sum_{j=1}^{n} \frac{1}{\lambda_j} w_{ij}^2$$

Hawkins proposed to flag any instance i that renders values for $T_2(X_i)$ inside a so called *outlier region* (Davies *et. al.* (1993)). López *et. al.* (1994) applied a closely related procedure also based upon PCA, to handle daily rain datasets. They proposed to flag an instance when for any one $j \in [K_1, K_2]$, the projection $w_j(X_i) \notin [LB_j, UB_j]$ being LB_j and UB_j lower and upper bounds which define the non-outlier region for projection $w_j(X_i)$. Those limits are derived from the distribution of $w_j(X)$. The eigenvalues themselves are not required as well as any specific distribution for the data.

This paper follows almost the same idea, but since we are now working with categorical data some details need to be discussed.

It should be pointed out that, even in numerical datasets, usually the mean value and the Principal Components are real vector values, and so are the projections of the dataset on the PC, which are called here scores. That holds even if the data are integer or even binary numbers. For example, in a rain dataset, all values are integer and positive, but the scores are real, i.e., they belong to a different number category. When considering categorical binary answers a similar situation arises. However, even real, the possible values are limited due to a combinatorial problem. We are implicitly requiring that this finite number is a large number (in the experiments, 2^{69}-1) and the reason is presented below.

Once the data (without missing values!) are binarized and presented in table (or matrix) format, the PCA can be performed straightforwardly. Principal components can be derived as the eigenvectors of the covariance matrix (Lebart *et. al.* (1977)). Let's call E the square nxn matrix whose columns are the eigenvectors, which satisfy $E\Sigma E^T = \Lambda$, being Σ the sample's covariance matrix, and Λ a diagonal matrix which holds the (sorted in ascending order) eigenvalues. "n" is not the number of controlled questions but the sum of all the options within them. It is assumed that the population is big enough to represent properly the true covariance with the sample´s covariance matrix.

Other subtle requirement should be stated: the procedure will not be of use if the number of options for the answers is low, because the distributions won´t look like those of continuous data. Notice that the real numbers w_{ij} are not arbitrary because they arise from a finite number of possible answers.

Anyway, since the matrix is range-defective due to the logical interrelationships between mutually exclusive answers, there will be some zero eigenvalues. This makes a slight difference with the situation for quantitative data (Hawkins (1974), López *et. al*. (1994)) where the Σ matrix is positive definite. The matrix of scores is defined here as:

$$W = E(X - \mu)$$

being X the binary data (one row for each record) and μ the arithmetic mean (among columns) of the matrix X. Matrix W has the same dimensions as matrix X, and its column-wise mean is zero. This is a linear transformation of the original data, and so each element w_{ij} depends directly upon *all* the elements x_{im}, where m ranges from 1 to n. This is an important fact, because the discrete distribution of the linear combination is completely different from the dichotomic one for x_{ij}, as it is shown in fig. 1.

Two facts should be remarked:

a) the sampled probability density function looks like the one of an ordinary continuous variable, even though it is based on a linear combination of dichotomic terms.

b) Its shape is different depending on the index of the score, following the same behavior noticed for scores derived from continuous variables, being more symmetrical as the index increases.

That's why we claim that the same procedures reported there could be used from now on. Once the sample distribution is created, confidence limits can be calculated. These values will define the outlier region (Davies *et. al*. (1993)) but without assuming any particular distribution shape. Why do we claim that this is the outlier region?. Fig. 2 shows the sampled probability distribution function for the given database of some of the scores and the arrows point to two values: those marked with an "o" correspond to the original answers for a particular record; those marked with an "x" are related to the same record, but now contaminated by modifying one of the answers. In this particular case, it was imposed that the house is equipped both with a color and a black and white TV set, while originally it has only black and white. Notice that the effect is important mostly in the "weakest" scores (i.e. those associated with the lower eigenvalues of matrix Σ) and that the ones associated with the "strongest" ones are only minimally modified. The proper limit between the "weakest" and the "strongest" is to be determined, and some guidance is given below.

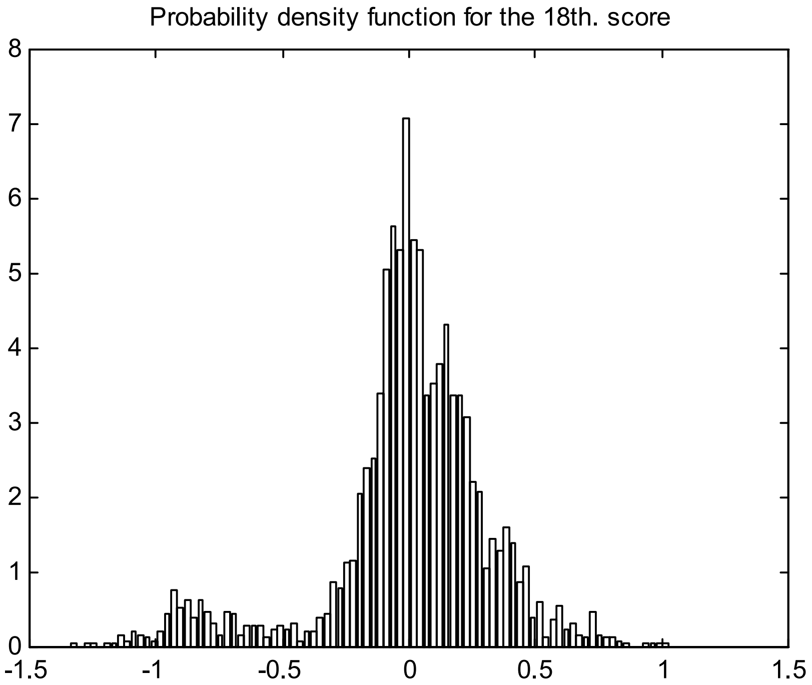

Figure 1 Example of the distribution of the 18th. score

A known fact is that each of those Principal Components associated with low eigenvalues have significant weights only with few variables in the original data. That implies that controlling the outlier region of one or two weak scores protects only some of the variables, which may be unadvisable. Those Principal Components associated with larger eigenvalues are typically insensitive regarding outliers, as it can be seen in fig. 2, so they should be avoided for our purposes. Summing up, neither too few or too many scores should be checked, and the appropriate number is a matter that is not uniquely solved in the literature.

Some rule of thumb suggest to neglect those terms whose associated eigenvalues are over a previously defined threshold. Hawkins (1974) suggests a more refined criteria, which chooses the limit in order to protect all the variables by proper inspection of the elements of matrix E . He did not formalize the criteria, so we will propose some objective one. The rows of matrix E are related with the original variables. Assume that K_1 is the index of the first non-zero eigenvalue, and K_2 is another integer index to be determined, ($K_1 < K_2$ because we assumed that the eigenvalues are sorted). In order to assure that at least once the variable X_j significantly affects some score, at least one of the eigenvectors with index ranging from K_1 to K_2 (i.e. columns in matrix E) should have a non negligible weight. The weights are the elements e_{jk} of row j of matrix E while $k \in [K_1,K_2]$, and they should be considered in absolute value for this purpose. The limit for negligible-not negligible is based upon a threshold value. If any abs(e_{jk}) is larger that such threshold for some $k \in [K_1,K_2]$, the variable is said to be *protected.* The threshold value cannot be chosen as a fixed constant like 0.17, because (due to normalization) the e_{jk} are related with the size n of matrix E . So the proper threshold should take this fact into account. Since a mathematically valid eigenvector could be $\left(1,1,1,\ldots,1,1\right)/\sqrt{n}$, we choose as a threshold value a multiple of $1/\sqrt{n}$, now independent of size n itself. In the simulations the chosen multiple was 0.15, and the resulting range $[K_1,K_2]$ was [21,45].

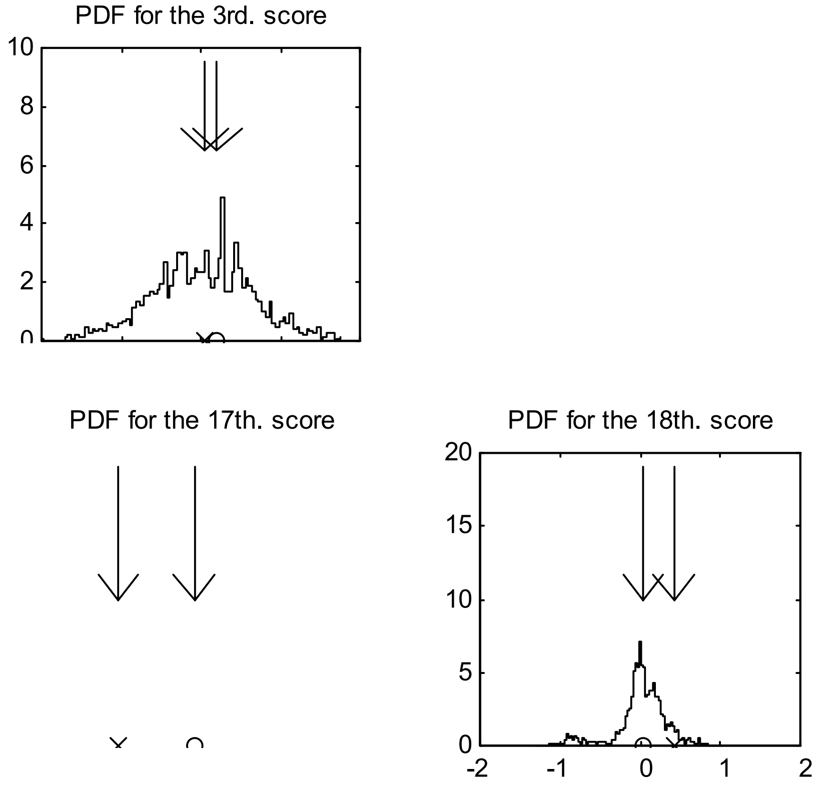

Figure 2 Example of the effect of a single outlier in a particular record

Once the limits K_1 and K_2 are defined, the sampled probability density function can be created for each score, and limits for the outlier region arise for each k, $k\in[K_1,K_2]$. The procedure is now straightforward, and it implies:

a) for each k-th score, look for records with values a_{ik} in the corresponding $k^{\underline{th}}$ outlier region, $k\in[K_1,K_2]$

b) once those records (rows) are retyped (and maybe modified or not), they can be included back in population X and new values for μ, E and outlier regions are calculated.

IMPROVEMENTS OVER THE DUPLICATE PERFORMANCE METHOD

The procedure is iterative, and some stop criteria should be given. In each step, the dataset is classified in two categories. The first one holds the records which are likely to have an error, and the second one holds the ones accepted. When such a decision is made, it is certainly possible to reject good quality as poor or classify nondefective items as defectives; then, the associated error is called Type I. When a decision is made to accept poor quality as good (classify defective items as nondefectives), the error is called Type II (Minton (1969)).

We will denote as *number of contaminated records found* the successfully identified records which belong to the candidates set. That set is suggested by the algorithm, and its size (the *number of candidates analyzed*) depends strongly on the parameters, as will be pointed out later. Its quotient is an estimate of the complement of the Type I error:

$$CE_I = 1 - E_I = \frac{\text{(number of contaminated surveys found)}}{\text{(number of candidates analized)}}$$

and it measures the rate of success looking from the point of view of the reviewing process. It should be noted that in a production environment CE_I can be measured by the end user without knowing the total number of errors (i.e., without retyping twice the whole dataset).

Another important number is the probability associated with a purely random choice, i.e. without using any rule in selecting the candidate set. As long as the procedure goes forward, an *accepted* set is created. The Type II error associated is defined by the quotient

$$E_{II} = \frac{\text{(number of contaminated surveys not found)}}{\text{(total number of "classified as acceptable" surveys)}}$$

This quotient can be expressed in more rigorous terms, as:

$$E_{II} = \frac{\begin{pmatrix} \text{initial number of} & - & \text{accumulated number of} \\ \text{contaminated surveys} & & \text{contaminated surveys found} \end{pmatrix}}{\text{(total number of surveys - accumulated number of candidates analized)}}$$

The E_{II} value also measures the probability to locate an error in the *acceptable* dataset with any blind (or random) procedure like the standard duplicate performance. Instead of presenting the evolution of the E_{II} index, a clearer measure of success is used, and it was defined as

$$\eta_2 = \frac{\text{(accumulated number of contaminated surveys found)}}{\text{(initial number of contaminated surveys)}}$$

This statistic monotonically increases from step to step, and it is bounded by 100 %, which implies that all the contaminated values have been located. It will also allow to compare directly the improvement over the standard duplicate performance method.

V.- Results

The calculations were carried out for 1, 2 and 3 wrong answers per record. Figure 3 shows the results for the first three steps in terms of the ratio η_2 for 100 replications of the experiment. The best results arise for a marginal value of 0.10%, where the methodology were able to locate 25% of the original errors (and in some cases, nearly 50%) *in a single step* of the procedure. The case of 0.01% looks very striking, because it represents two different bimodal, bell shaped distributions. This behavior is connected to some extent to the small number of records involved, as it will be shown later, and the picture is still incomplete because it does not show the effort involved in each step. Again, the value chosen for the marginal value is not crucial.

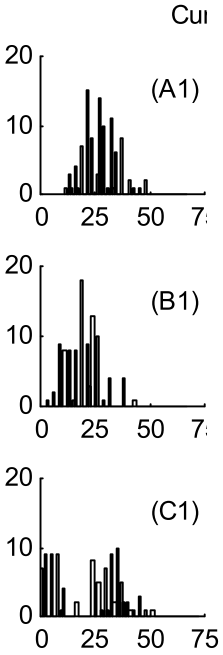

Figure 3 Distribution of the accumulated fraction of the total errors located up to the first three steps. Plots derived after 100 experiments, modifying 3% of the records with 2 errors each.

Figure 4 is itself a global summary of the behavior of the method. The x-axis is the fraction of the total dataset retyped, while the y-axis represent η_2, the fraction of the total errors found. We should emphasize that the continuous line indicates the locus of the theoretical evolution of the standard (blind) duplicate performance method, i.e.: by typing the x% of the whole dataset, the same x% of the errors were removed (notice that the line goes through the (20%,20%) point). For *any* choice of the marginal value, the methodology proves to be better than the standard duplicate method, and since the behavior was very similar, only the case of 0.10% is shown. The dotted line is the best you can attain: retype first only those records that have errors. In the figure while retyping only 5% of the original data (x-axis) we can locate an amount of the original errors ranging from 25-60%, and when retyping 10%, 40-75% can be located.

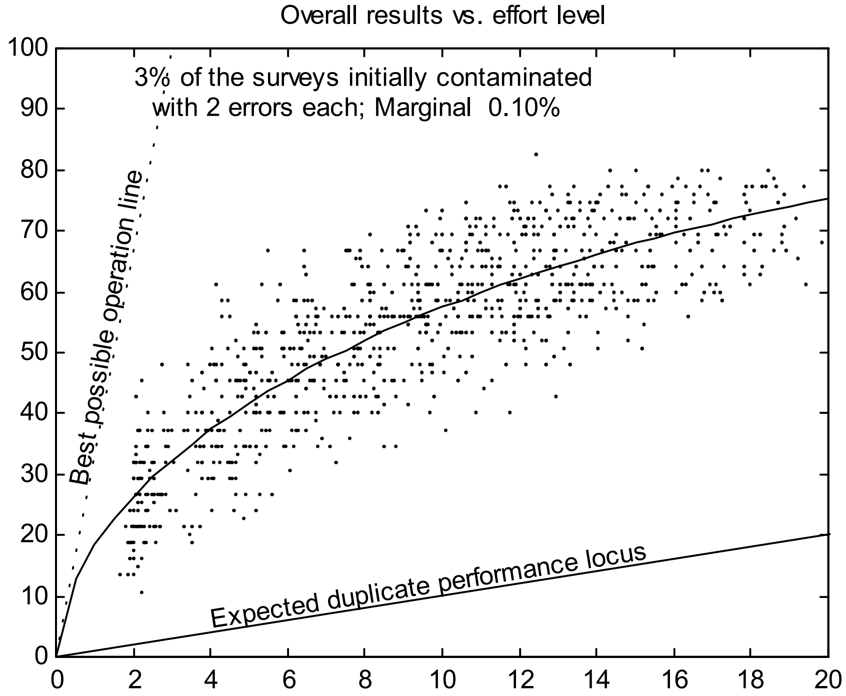

Figure 4 Evolution of the remaining errors against the retyping effort for the suggested depuration order and the blind retyping. Plots derived after 100 experiments, modifying 3% of the records with 2 errors each.

Further work will show a degraded performance, because the “worst” errors have already been located. The limit goal of the procedure will be also the (100%,100%) point, because if all the data are checked we assume that all the errors will be removed. This procedure is intended to be applied for *partial* retyping.

The previous figures have presented the results with the records contaminated with 2 errors each. As expected, with 3 or more errors per record the results will be better while with 1 error they will be worse. For the 3 errors per record case, figure 5 show that after retyping 10% of the database, 50-85% of the errors have been corrected. For the case of a single error per record, fig. 6 shows that after retyping 10% of the database, only 25-50% of the errors have been corrected at most. The reader may notice that the cloud do not show any point over 16%; that's because we limit ourselves to 10 steps in the procedure. Even in this difficult case, the method is typically 4 times better than the blind retyping.

IMPROVEMENTS OVER THE DUPLICATE PERFORMANCE METHOD

Some results regarding the initial number of erroneous records (not presented) show that the behavior of the best fit curve is almost independent of such value, but the dispersion is lower for larger initial number of erroneous records. This fact is a very desirable property, because poor quality dataset can be handled without loosing performance.

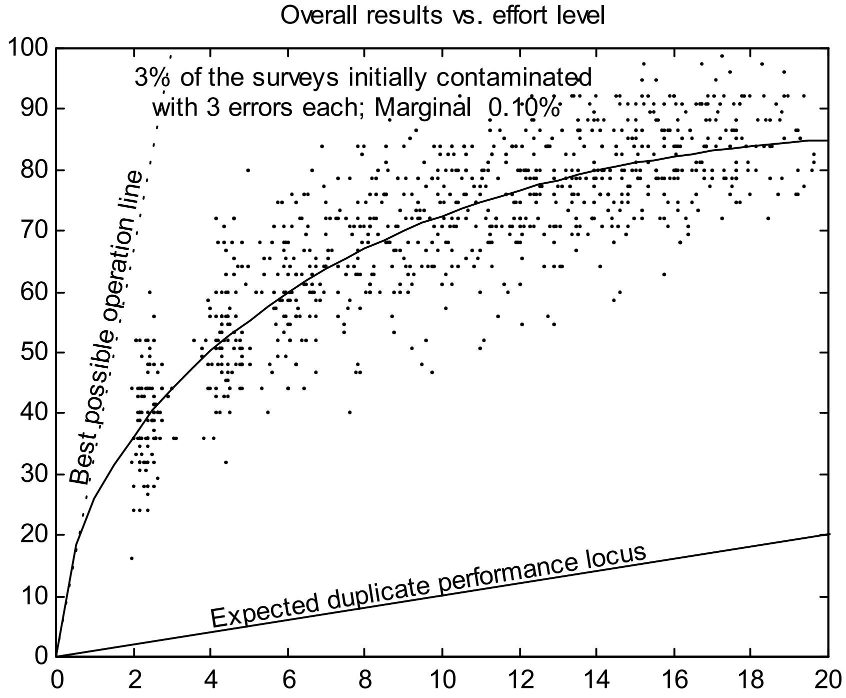

Figure 5 Evolution of the remaining errors against the retyping effort for the suggested depuration order and the blind retyping. Plots derived after 100 experiments, modifying 3% of the records with 3 errors each.

VI.- Discussion

Comparing the use of logical edits against the present methodology, some clear differences arise. The methodology proposed in this work does not require any expert, since the “rules” (if any) are embedded in the population. Even the dichotomic answers (like marital status, sex, etc.) which are mutually exclusive, are handled gracefully, and need not to be analyzed separately. Moreover, when the population is updated using mostly the same questions, but with changes in some of them, all related rules should be revised. If a question is ambiguous, the rule can be wrong, while the proposed methodology probably will flag the answers as “uncorrelated” and will remove automatically from the feasible set.

Since the mere retyping is a completely “blind” methodology, it will locate equally well errors in “unusual” as well as “typical” individuals, keeping the variability of the dataset, while both the proposed methodology and the logical edits are oriented toward flagging only those errors which make a particular individual “unusual”. Then they will diminish the variability of the dataset.

However, it should be admitted that the application either of logical rules or mere retyping do not require a large population of individuals, while this methodology implicitly does. Other limitation of the reported methodology is that not all the questions can be controlled, either because of almost trivial answer or low correlation with other answers. Moreover, it cannot handle individuals with missing values.

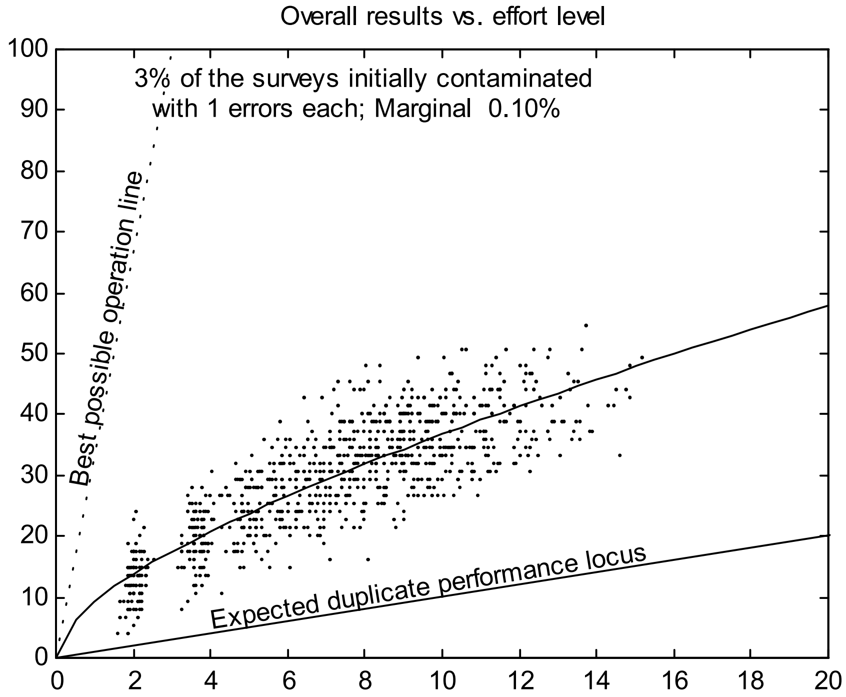

Figure 6 Evolution of the remaining errors against the retyping effort for the suggested depuration order and the blind retyping. Plots derived after 100 experiments, modifying 3% of the records with 1 errors each.

The numerical procedure is quite simple. It requires first to transform all answers to a “check box” format, so only ones or zeros will be admitted as answer. Then, the covariance matrix is constructed and its eigenvectors calculated, and a new table of projections (scores) of the original individuals over the eigenvectors is created. By analyzing the eigenvectors, a critical set of the scores is chosen in order to calculate for each an outlier region. Every individual with at least one of its scores lying on those region should be retyped. All the procedure can be automated. Once calculated the eigenvectors and the critical set, it can be applied even during the first typing process, allowing for near real time quality control.

The sensitivity to some parameters have been tested during the work, and for others not. Among the first, the margin (related with the number of individuals to be retyped in each step) was only weakly significant. The methodology for selecting the principal components to check seems feasible, but no further tests have been done.

For perfectly uncorrelated answers as a limiting case, the procedure is equivalent to looking for answers with low probability, which is also a feasible procedure.

IX.- Conclusions

The problem of quality control of categorical data is treated with a methodology derived from statistical procedures for quantitative data. Two other alternatives can be analyzed, the duplicate performance method and the use of logical edits. The first is very simple and popular, and requires typing again the same dataset. Its ability in locating errors for a given typing effort is known to be low. The use of logical edits strongly rely on the existence of an expert, which should prepare a set of rules, expressed in terms of logical relationships between the answers. When any of them is not met, the record is flagged as unusual, and retyping is performed. Here an alternative is proposed in order to reorder carefully what should be retyped.

Some limitations of this procedure are: a somewhat large (yet undefined) population is required as well as a minimum number of options for the answers, it cannot handle missing data, and depending on the inherent characteristics of the population, some answers or options for answers are not checked. The users for a methodology like this are still those which are either collecting or using the raw data; we are not giving any tool to check derived statistics (like averages in a region, etc.).

X.- Acknowledgments

The author is indebted to the authorities of the National Bureau of Statistics of Uruguay, which allowed access to the raw data used in this work, and to Prof. Friedrich Quiel, Royal Institute of Technology, Stockholm, who provided helpful discussion and suggestions while preparing the manuscript.

XI.-References

Coale, A.J. and Stephan, F.F (1962) The case of the Indians and the teen-age widows, *Journal of the American Statistical Association*, Vol. 57, No. 298, 338-347

Davies, L. and Gather, U. (1993) The identification of multiple outliers. *Journal of the American Statistical Association,* Vol. 88, No. 423, 782-801

Fellegi, P. and Holt, D. (1976) A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, Vol. 71, No. 353, 17-35

Minton, G. (1969) Inspection and correction error in data processing. *Journal of the American Statistical Association*, Vol. 64, No. 328, 1256-1275

Hawkins, D. M. (1974) The detection of errors in multivariate data using principal components. *Journal of the American Statistical Association*, Vol. 69, No. 346, 340 -344

Little, R. J. A. and Smith, P. J. (1987) Editing and Imputation for quantitative survey data*. Journal of the American Statistical Association*, Vol. 82, No. 397, 58-68

López, C.; González, E. and Goyret, J. (1994) Análisis por componentes principales de datos pluviométricos: a) aplicación a la detección de datos anómalos (in spanish) Estadística, Vol. 6, No. 146-147, 55-83. (English version available at http://www.fing.edu.uy/cecal/reports/rep92_1/papera.html)

Paradice, D.B. and Fuerst, W.L. (1991) An MIS data quality methodology based on optimal error detection. *Journal of Information Systems*, Vol. 5, No. 1, 48-66

IMPROVEMENTS OVER THE DUPLICATE PERFORMANCE METHOD

Strayhorn, J. M. (1990) Estimating the errors remaining in a data set: techniques for quality control, *The American Statistician*, Vol. 44, No. 1, 14-18

Appendix 7

López, C. and Kaplan, E., 1997, "A general purpose procedure for locating outliers in multivariate time series: Application to an hourly wind dataset"

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.

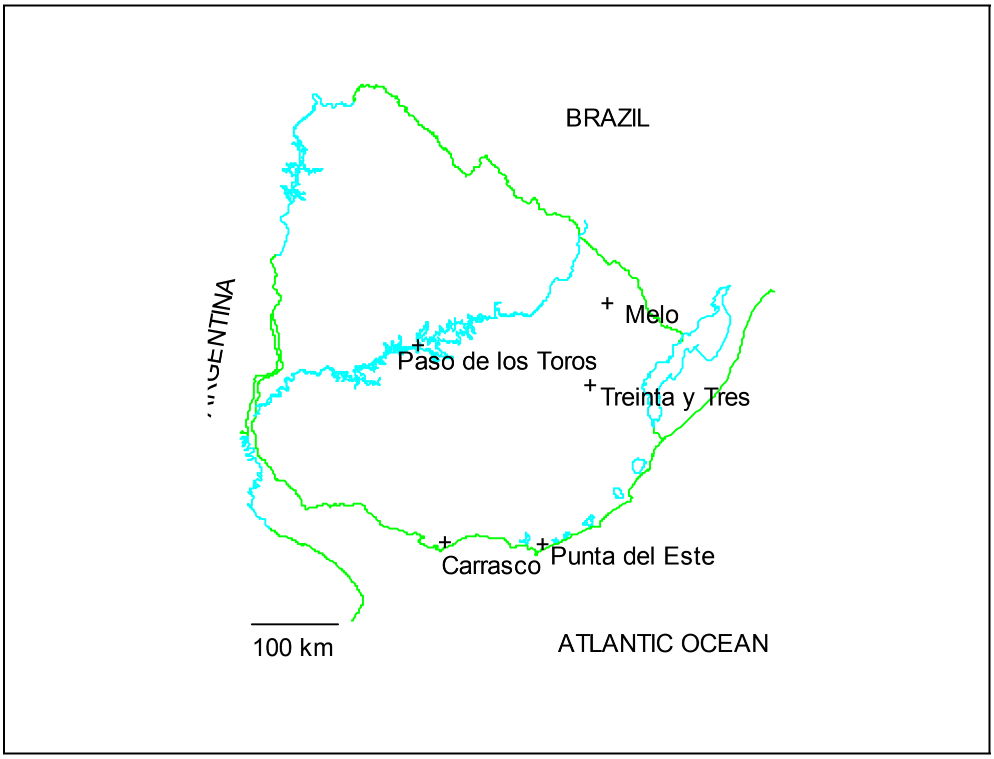

Figure 1 Location of the weather stations

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.

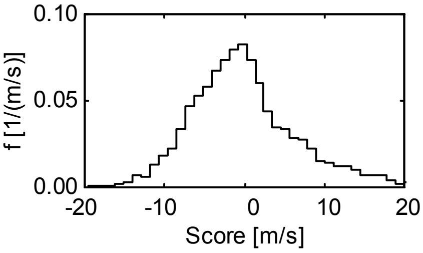

Figure 2 Sketch of some properties of the scores 1 (left) and 2 (right). On top the probability density function; on the middle the power spectrum and on bottom the self correlation against the lag.

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).

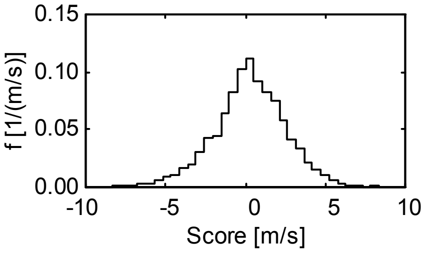

Figure 3 Sketch of some properties of the scores 3 (left) and 4 (right) as presented in fig. 2. Notice the change in scale for the power spectrum. The peak near 0.04 is due to daily variations.

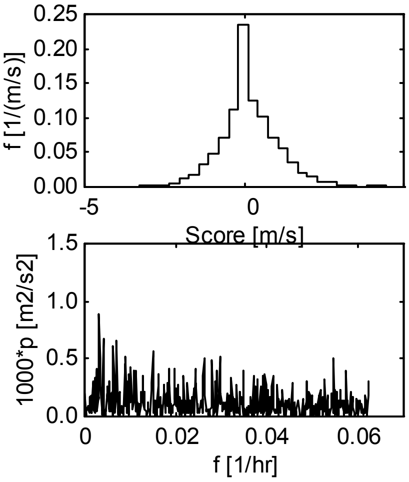

Figure 4 Sketch of some properties of the scores 9 (left) and 10 (right) as presented in fig. 2. Notice again the change in scale for the power spectrum, and also in the x-scale for the pdf.

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.

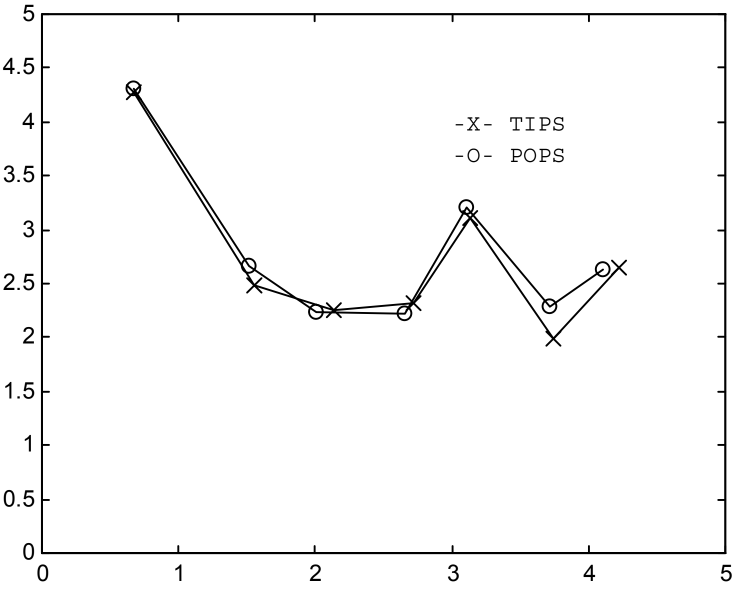

Figure 5 Experimental values of the RMS obtained while depurating the original database. X-axis stands for the fraction of the total database already checked. Y-axis stands for the RMS of difference between wrong values and correct ones on paper.

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.

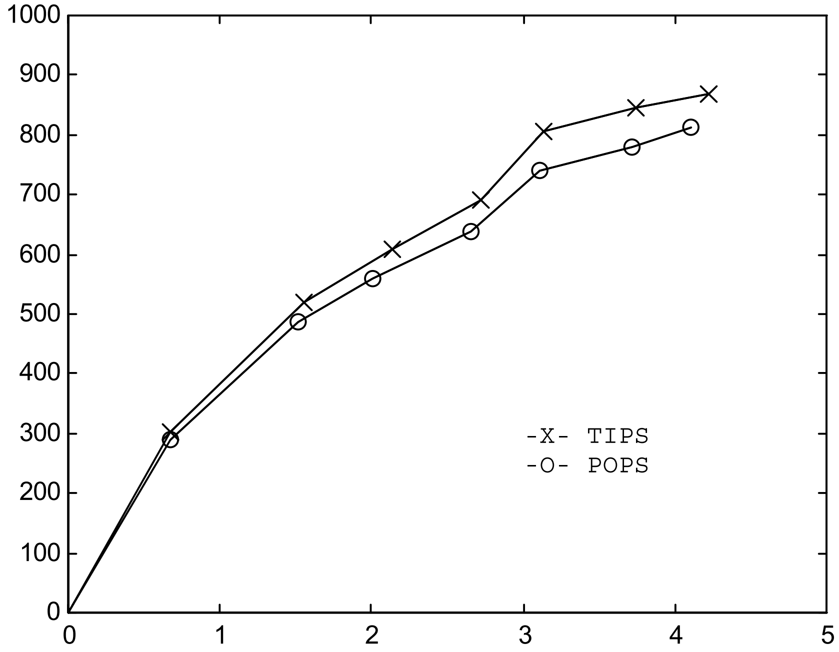

Figure 6 Total number of errors found for a given effort..

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.

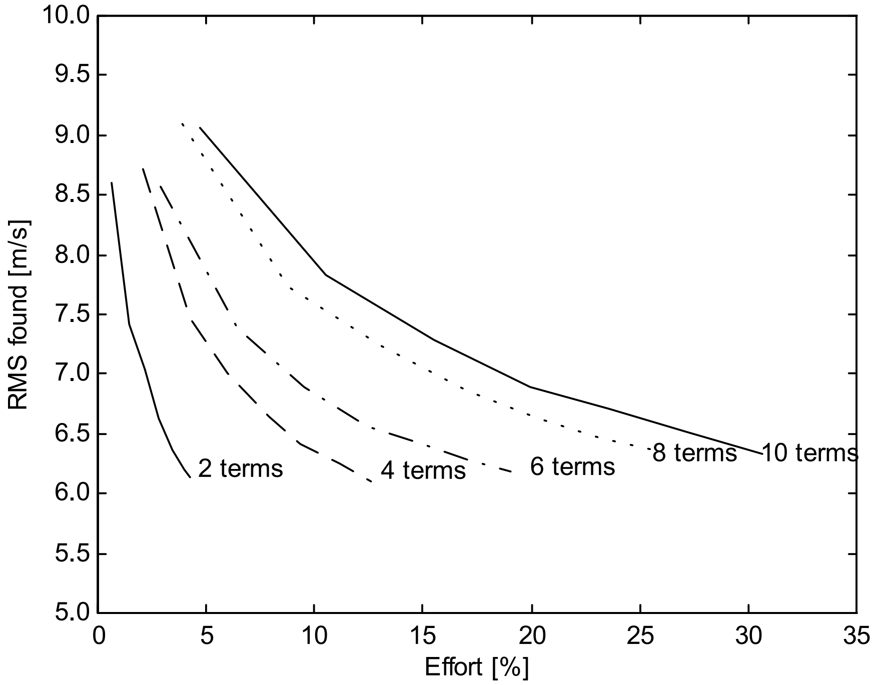

Figure 8 Evolution of the RMS found for the same initial noisy dataset in terms of the effort

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.

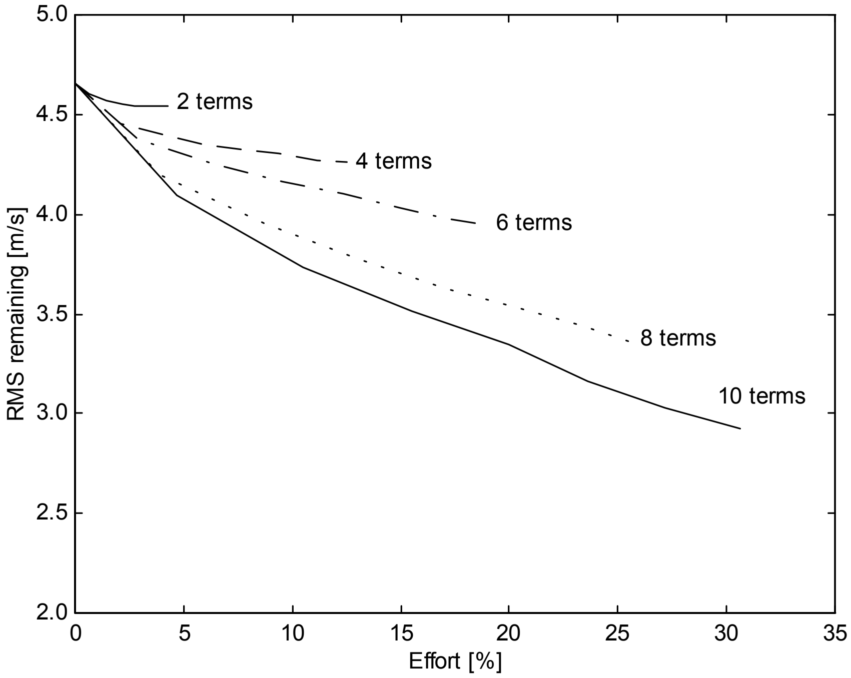

Figure 8 Simulated evolution of the remaining RMS vs. the effort

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.

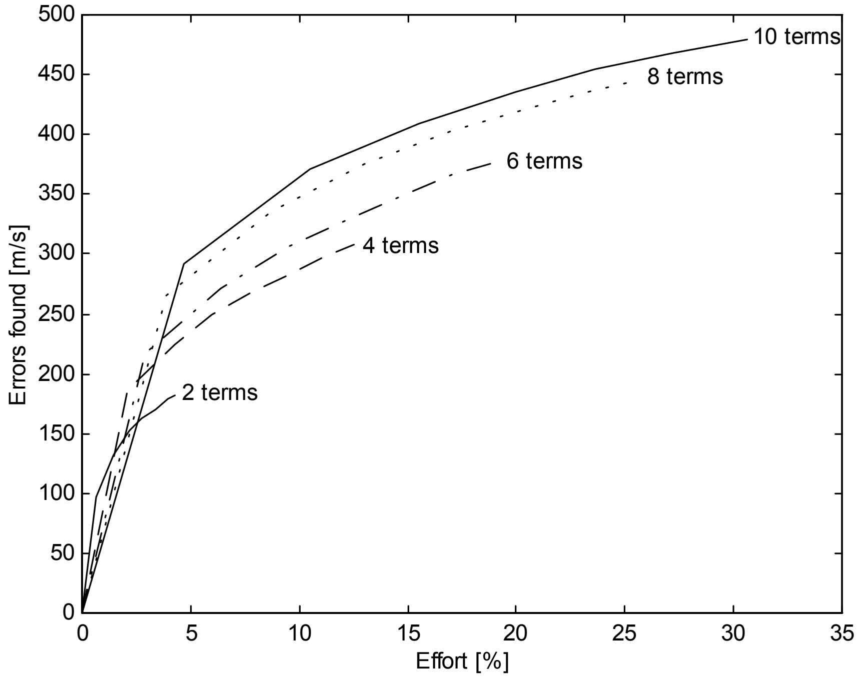

Figure 9 Simulated evolution of the total number of errors found vs. the effort

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)

Appendix 8

López, C. and Kaplan, E., 1997, "A new technique for imputation of multivariate time series: application to an hourly wind dataset"

A new technique for imputation of 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.

The outlier detection phase is presented in a companion paper, and here the different techniques applied in order to imputate the missing values are described. The comparative results obtained with an hourly dataset of 15 years long are also presented. Two different problems have been simulated numerically: systematic missing values (i.e. at fixed hours) and non systematic ones.

Five different criteria were applied: imputation with the historical mean value; linear time interpolation within single station records; optimum interpolation (kriging) and the two newly developed **P**enalty **O**f the **P**rincipal **S**cores and linear **T**ime **I**nterpolation of the **P**rincipal **S**cores which considers all station records in a multivariate fashion; they prove to be the most accurate for this particular wind dataset. There is also some evidence of oversampling in time.

1. Introduction

1.1 Presentation of the problem

Since 1988 the Team was involved in evaluating the National Wind Energy Resource. Although it was not an explicit objective of the project, it was necessary to complete the time series to the longest available period. With these goals some algorithms have been implemented for imputation of missing values in the series. Some test have been carried out that confirm the kindness of their performance in controlled cases.

Three different methods have been tested: a) Time interpolation of the principal scores (TIPS) (includes standard time series interpolation as special case) b) Penalty of the principal scores (POPS) and c) Optimum interpolation (Gandin, 1965).

The first two were developed in López *et al.* (1994b). The third is a standard interpolation procedure in the inicialization of mathematical models in meteorology (Johnson, 1982) that allows to find estimates not only in the stations of measuring, but also in another points of the domain of work. In this case their performance has been analyzed in the case in that the point to interpolate is one of the measuring stations. In the work of López *et al.* (1994b) on rain data, the TIPC resulted to be a poor method, while for the case of wind it was the best one.

1.2 Methodological background

Objective analysis methods are very common in meteorology (see Haagenson, 1982, Johnson, 1982, etc.) since they have been designed to produce an interpolated field using only data observed in irregularly distributed networks. This situation gives, on principle, a way to overcome the problem of missing values, because they can be estimated using available information.

For some applications the missing values are not a problem (for example, *find the extreme values, calculate an annual average*, etc.) while for others they are critical. The existence of well established procedures led in the past to a marginal interest for the problem, clearly observed in the scarcity of specific work found in the literature. We believe that, in most cases, missing values are simply ignored, under the implicit hypothesis that those errors appear at random. Such hypothesis is rarely tested nor verified.

Missing values are extremely important in statistics and social sciences, where even books on the topic can be found (Rubin, 1987) mentioned results from international working groups. In such areas there exist both crude and sophisticated imputation methods. For example, the one suggested by the U.S. Census Bureau (Rubin, 1987) assigns a randomly chosen value among those events which other values coincide, or are below a so defined "distance" from the target one.

Another simple method is to make a linear regression using available data. Usually such model is build up least squares criteria, principal component analysis, etc. (Stone et al. 1990) All of the above produce for each missing value, a single candidate. Following Rubin, 1987 "..it is intuitively clear that by imposing an "optimum" value, the variability will be underestimated". The author suggests that more than one candidate might be produced, and he described the techniques typically used for surveys. The general idea is to create, for each missing value, a small number m of candidates, and consider that you have m datasets. The method is workable if there are a low number of missing values; its results are usefull, but require more computing time and also more space (to store the multiple imputed values). We refer to Rubin (1987) for further details.

2 Testing Methodologies

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 National Department of Energy (DNE) Wind Energy Program. This also conditions the periods to work with, including part of the years 1990-1991 and the year 1984.

The work carried out consisted in:

a) removing temporarily those values to be imputed

b) for each method

  b.1) eliminating all missing values

  b.2) calculating RMS and mean between new and true values

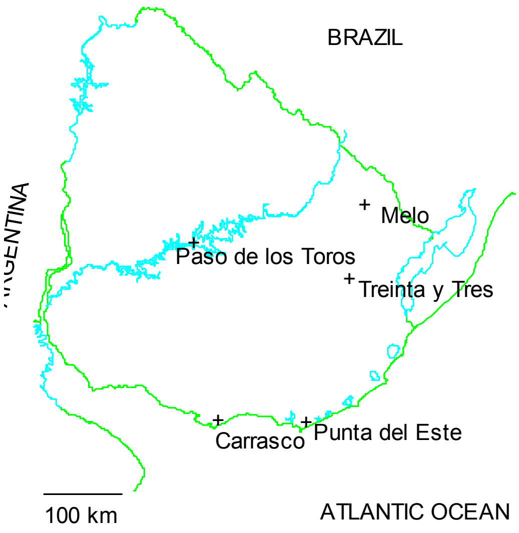

Figure 1 Location of the weather stations

The methods were presented in full in López *et al.* (1994a and 1994b) applied to the case of pluviometric data (daily values). A synthesis will be introduced pointing out the differences among rain vs. wind here managed. 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 some absence in them. The quality control process requires either measured or estimated data for the considerate event, in all the stations, so an imputation for the missing values is required.

The Principal Components Analysis (PCA) was already used for the same wind dataset in Cisa *et al.* (1990). 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 in particular. Any complete event has associated 2n measurements (components or 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 the principal scores are the projection of the space in that base (Lebart *et al.* 1977). Since both bases relates each other by means of a linear relationship, it is equivalent to manage the temporal series of the original measurements in the stations or the temporal series of the principal coefficents.

The time series of the principal coefficents 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 figures 2 to 4 some of the distributions of the observed coefficents are introduced as well as its power spectrum.

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 Cisa *et al.* 1990. From the analysis of the figures 2 to 3 is deduced that the series of the corresponding scores 1 to 3 varies smoothly, in opposition to the registered in the subsequent figures. The spectrum shows a noisy pattern, and the selfcorrelation decreases more sharply with the lag. It can be that, in the case of existing any missing values, it would be reasonable to perform a linear interpolation in time for the scores of minor index. The other scores are typically of minor or greatly minor importance (compare the dispersion of the figure 4 with the one of the figure 2) and they can be neglected. So after interpolation of the main scores and setting to zero the remaining ones the complete set of scores can be obtained for the event with missing values. By means of the linear mentioned transformation the tentative registrations are calculated, but only those that were lacking are incorporated. A more precise estimate of the interpolated scores is now possible upon incorporating the values indeed measured corresponding to the event. For more detail, the reader refers to López *et al.* 1994a, 1994b.

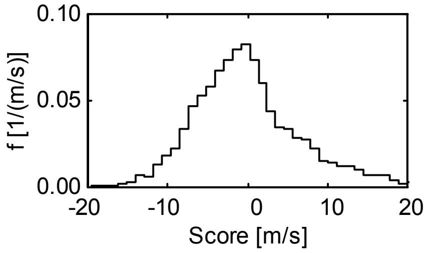

Figure 2 Sketch of some properties of the scores 1(left) and 2(right). On top the probability density function; on the middle the power spectrum and on bottom the self correlation against the lag.

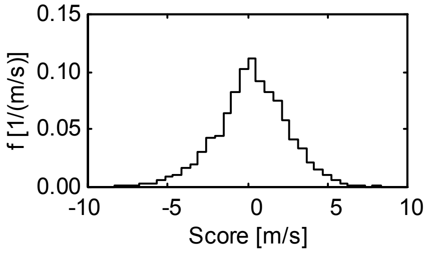

Figure 3 Sketch of some properties of the scores 3 (left) and 4 (right) as presented in fig. 2. Notice the change in scale for the power spectrum. The peak near 0.04 is due to daily variations.

This procedure of temporal interpolation will be named Time interpolation of the principal scores (TIPS) hereinafter. As a particular case, the standard linear interpolation between registers of the same stations is an special case, when all the principal scores are included in the interpolation. The method

A new technique for imputation of multivariate time series of Penalty of the principal scores (POPS) was also used, as introduced in López *et al.* 1994b. It is based upon the fact that the weakest patterns have associated usually very small scores; so the functional (suggested for the first time by Hawkins, 1974)

$$\mathrm{S}(u_1, v_1, u_2, v_2, \ldots, u_n, v_n) = \sum_{\mathrm{i} \in \mathrm{p}} \frac{\mathrm{a}_\mathrm{i}^2}{\mathrm{w}_\mathrm{i}} \qquad \mathrm{i} = k..2n$$

will be also small. a_i is the i-th score, and w_i is a weighting score. The key idea is that any missing value u_j or v_j can be estimated directly by minimizing the S functional.

$$\frac{\partial \mathrm{S}}{\partial u_\mathrm{j}} = \frac{\partial S}{\partial v_\mathrm{j}} = 0 \qquad \mathrm{j} \in r$$

where u_j, v_j denotes the missing component u and v in the j-th station and being r the set of stations with missing values for the event. The system of equations is linear in u_j, v_j and of moderate size. It may happend that the optimum values still are unacceptable, by application of the criteria documented in López *et al.* 1994a, but such constraint has not been imposed in our experiments. The particular case k=1 corresponds to minimize the Mahalanobis distance to the mean value.

As a final point, it should be stressed that this procedure neglects the information of the temporal self correlation of the coefficents. For details, again the reader refers to López *et al.* 1994b.

3 Results

Two experiments were carried out, depending on the systematic nature or not of the missing values. The dataset corresponds to the years 1990-91 after removing errors

3.1 Systematic location of missing values

Typically in the northern zone of the country most stations take readings at 8, 14 and 20 hours, as some stations from Southern Brazil and the Argentina do. Only the stations of Artigas, Rivera, Salto and Paso de los Toros in Northern Uruguay take hourly wind values. It has been evaluated the possibility of imputate such systematic holes. In order to have a frame with what compare, we have analyzed again the five stations of the southern zone: four of them were considered with data only at 8, 14 and 20 hours local time, while Melo was taken as fully hourly. The three above mentioned methods were applied to this case, being the results presented in Table 1.

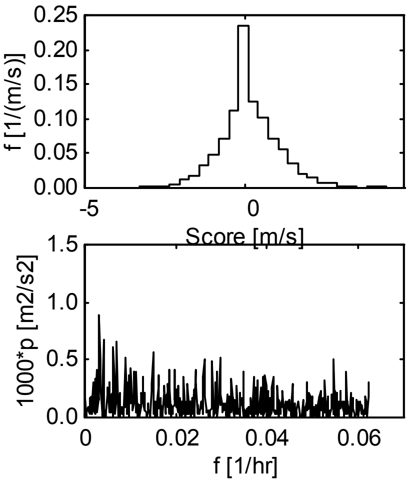

Figure 4 Sketch of some properties of the scores 9 (left) and 10 (right) as presented in fig. 2. Notice again the change in scale for the power spectrum, and also in the x-scale for the pdf.

As a reference, it has been evaluated the error when every component x_j was imputed simply with its mean value for the period, and by application of the Optimum interpolation, which renders in this case for each hour, the same value for every station. The outputs are consistent in that the TIPC performs the better, without a clear difference between interpolating the 4 first scores or all of them. The experiment simulates a population of 60924 absences in 4 of the 5 considerate stations, leaving only the hours 8, 14 20 like well-known data, being the total usable population in the calculation of the eigenvectors of 83724 values (corresponding to the hours with mensurements in all the stations). The critera adopted imputated the missing values with a mean error of 0.10 m/s and a RMS of 2.05 m/s, calculated in relation with the original data (Table 1). The procedures were also evaluated in that the absences are at random. The results also support as a suitable choice interpolating all of the 10 coefficents, and will be presented in the next section

3.2 Non systematic location of missing values

15197 absences in the population of 83728 were created at random dates and stations, being approximately 20% of the total. In this case the selected criterion of interpolate with all the terms (terms 1:10) renders a mean error of 0.03 m/s and RMS of 1.67 m/s between the calculated data and the original values (Table 2).

The results from Table 2 are valid for a *single* random set. Other independent runs revealed that in some cases a minimum in k = 7 could be noticed, more in accordance with the previous results for rain. However, the corresponding optimum RMS error was not very different from the value for k = 10. The objective (goal) function was the RMS of the population of differences between the calculated value and the one indeed measured. The optimum value is approximately 2 m/s, which is acceptable for wind speed. It should be noticed that interpolating with all 10 terms is equivalent to an independent interpolation of each station's time series.

4 Conclusions

From the performed experiments it could be inferred that:

a) The algorithms employed in order to imputate values performed very satisfactorily. They outperform the standard procedures.

b) The near optimum value obtained using the standard interpolation of the time series suggests that, at least in Uruguay, the wind records are oversampled in time. When an artificial undersample is introduced the optimum number of interpolated terms decreases. We provided a physical explanation of the phenomena based upon the spectrum of the score's time series.

Table 1 Calculation of the root mean square (RMS) and the mean of the error (measured value - calculated value) upon imputation assuming complete Melo and the another 4 stations with registrations in the hours 8, 14 20 only. Data from the year 1990-91. In boldface the most significant outputs.

Interpolation			Penalization		
Interp. terms	RMS (m/s)	Mean (m/s)	Penalized terms	RMS (m/s)	Mean (m/s)
1:10	2.06	0.09625	10:10	3.41	0.10094
1:9	2.06	0.09669	9:10	3.41	0.10274
1:8	2.05	0.09613	8:10	3.39	0.10452
1:7	2.06	0.09671	7:10	3.28	0.06608
1:6	***2.05***	***0.08151***	6:10	3.26	0.06191
1:5	2.06	0.09585	5:10	3.23	0.04485
1:4	2.05	0.09541	4:10	3.21	0.01852
1:3	2.05	0.08414	3:10	3.40	0.01686
1:2	2.11	0.08763	2:10	2.97	0.00177
1:1	2.73	0.07331	***1:10***	***2.84***	***0.05171***

Results obtained assigning the mean value	3.24	0.28839
Results obtained with the Gandin´s method	2.84	0.05353

Table 2 Calculation of the root mean square (RMS) and the mean of the error (measured value - calculated value) upon imputation assuming 20% of the data with missing values.
Data from the year 1990-91. In boldface the most significant outputs.

Interpolation			Penalization		
Interp. terms	RMS (m/s)	Mean (m/s)	Penalized terms	RMS (m/s)	Mean (m/s)
1:10	**1.67**	***0.03193***	10:10	6.83	0.19072
1:9	1.67	0.03092	9:10	7.78	0.09796
1:8	1.68	0.03401	8:10	7.99	0.13297
1:7	1.68	0.03286	7:10	8.27	0.14827
1:6	1.70	0.04416	6:10	7.02	0.02573
1:5	1.73	0.04740	5:10	5.34	0.03002
1:4	1.76	0.02857	4:10	3.33	0.10016
1:3	1.79	0.03594	3:10	2.73	0.06495
1:2	1.89	0.03005	2:10	2.35	0.04825
1:1	2.57	0.00417	***1:10***	**2.33**	***0.07813***

Results obtained assigning the mean value	2.76	0.03141
Results obtained with the Gandin´s method	2.37	0.07596

5 References

Cisa, A.; Guarga, R.; Briozzo, C.; López, C.; Alonso, J; Cataldo, J.; Canetti, R.; Acosta, A.; Penza, E.; Xavier, V.;Tozzo, A.; Estrada, J.; Bevc, A.; Maggiolo, G.; Chaer, R.; Rosenblatt, R.; Lamas, R.; Martínez, F. y Cabrera, R., 1990. "Proyecto de Evaluación del Potencial Eólico Nacional: Informe Final" Facultad de Ingeniería, Instituto de Mecánica de los Fluídos e Ingeniería Ambiental e Instituto de Ingeniería Eléctrica, Montevideo, Uruguay. 1000 pp. (in spanish)

Gandin, L. M., 1965. "Objective analysis of Meteorological Fields". Israel Program for Scientific Translations, 242 pp.

Haagenson, P.L, 1982. "Review and evaluation of methods for objective analysis of meteorological variables" Papers in Meteorological Research, V 5, N 2, 113-133.

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.

Johnson, G.T. 1982. "Climatological Interpolation Functions for Mesoscale Wind Fields". Journal of Applied Meteorology, V 21, N 8, 1130-1136.

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.

Rubin, D. B., 1987. "Multiple imputation for nonresponse in surveys". John Wiley and Sons, 253 pp.

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" Final report (in spanish). Part 2, Cap. 4. 83 pp.

Stone, M.; Brooks, R.J., 1990: "Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression" J. R. Statist. Soc. B, 52, N 2, pp 237-269.

Appendix 9

López, C., 1997, "An error model for daily rain records"

AN ERROR MODEL FOR DAILY RAIN RECORDS[1]

CARLOS LÓPEZ
Centro de Cálculo, Facultad de Ingeniería (11), CC 30
Internet: carlos@fing.edu.uy; http://www.fing.edu.uy/~carlos
Montevideo, URUGUAY

Abstract

Removing outliers from records prior of its use is a major concern in any technical or scientific field. Meteorology is not an exception, and an important effort in devise methods has been made to locate them despite the fact that it has been misconsidered as a purely technical task. The currently applied methods are very crude because they are mostly computerized versions of traditional criteria, failing to exploit the capabilities of modern computer systems. Extensive comparison among methods have not been done, no reliable statistical comparison among different outlier detection strategies can be made without a tool for generate instances of a database contaminated with artificial errors. This paper describes a heuristic model suitable to simulate the usual errors observed in a 30 years, ten stations, daily rain dataset, which has been carefully checked against typing errors. We will restrict ourselves to simulate only such errors. Some methods are discussed, namely: a) choosing at random other value in the same dataset b) choose at random other value for the same station c) model imperfectly some driving mechanism for the errors. The results will be compared with the observed problem, and from them we were able to show that options a) and b) underpredicts the difference between errors and true values, while even imperfect, option c) renders satisfactory results.

1. Introduction

Most of the literature related to observation errors in meteorology are devoted to systematic and random errors before the data is put on paper. Typical problems is representativeness of the measurements in relation to the surroundings, and measuring device problems.

A suitable error model (i.e. an algorithm) should address two problems: a) provide rules to select a date and a station candidate to be in error and b) assign an outlier for it. Since we are considering existing records, we are not concerned with systematic errors: the outlier should suffer from them as well. The literature on random errors in meteorology analyzed mostly the distribution of the differences between *truth* and *observation*, solving to some extent the phase b). However, typical analysis make comparisons between instruments of different accuracy, etc. which might not be the most important source of concern here, because we attempt to handle errors produced after the observation is put on paper for the first time. We will now review some previous work disregarding the particular meteorological variable considered.

One standard and possible approach suitable for random errors is to compare the observations against the outputs of a conceptual model of the phenomena (if it exists!). See for example Francis (1986); Hollingsworth *et al.* (1986). The latter reported that for the hourly wind, the differences between observations and predictions follow a gaussian distribution with zero mean and σ standard deviation (shortly N(m, σ^2)). However since he analyzed the operation of automatic equipment, he did not report any rule to assess *when* the errors are more prone to appear. Anyway, such procedure is suitable for only a limited number of meteorological variables, since a different and specialized mathematical model is required for each one. Requirements in computer time are also important, as well as availability on other data for use within the model.

[1] Presented at the 10th Brazilian Meteorological Conference. Brasilia, Brazil 26-30 October, 1998

The strategy of compare the output of a mathematical model with the observed values is common at Global Data Assimilation Centers (Gandin 1988; Di Mego 1988; Parrish 1992). Gandin (1988), asserts that the most significant part of the errors detected are those collected in developing countries, partly due to the fact that there are certainly less resources affected to data collection than those affected in developed countries. Unfortunately no reference to patterns of occurrences are reported.

For the case of homogeneous (i.e. same measurement units) variables one particularly simple model is the linear one, which is a first order approximation of the *parameter(x,y)* function. Hawkins (1993), declares that true errors are traditionally been assumed to be distributed as $N(0,\sigma^2)$. However, regarding the outlier identification problem and the robust regression one there is concern about the correctness of this part of the model, specifically that the error may come from a distribution with heavier tails than the normal. In his paper he analyzed some well known datasets, but none of them are of meteorological variables. Rocke and Woodruff (1996) demonstrated that in order to detect outliers assuming multivariate normal distribution the most difficult case is the one with shifting outliers but the same covariance matrix.

To some extent the task of modeling the pattern of missing values occurrences can be considered as similar to the objective a) of this problem. For example, Little (1992) addressed the problem in the regression framework. He differentiates among four mechanisms: 1) the missing value is independent on the data values 2) depends only on the value of X_1 for that case 3) depends on X_2, X_3, .. X_n for that case or 4) depends on X_2, X_3, .. X_n and Y for that case. We found that this problem is strongly dependent on the available data for the day, so mechanism 4) was finally adopted.

Reek *et al.* (1992) reported results using a climatological database collected through a cooperative network. The database has over 100 years long, an has been collected by volunteer weather observers. The database was progressively transferred to punch cards in order to help in prepare publications. The storage media started to be the magnetic tape in the late 1960s. In our opinion this datasets has many similarities with the one which will be described, and both the collection and typing procedure might be considered as representative for other countries as well. They describe also the set of rules which allow both for recognition and correction of erroneous data. The rules are predetermined, systematic, and empirically derived. Some of them might apply only to *that* dataset, but illustrates common situations.

As a final remark, we want to quote Gandin (1988). He said *"...for a long time, however, the Quality Control development was considered as a purely technical task, and no further investigations on rough errors were performed..."*. That explains why the topic addressed in this paper has received little or very few attention.

2. The available dataset

The description to be presented has been taken from Bidegain and Fontana (1996 personal communication). We have used data from 1960 to 1990 taken in the Santa Lucía catchment area, of 13600 km^2, located in the Southern Uruguay approximately between 55° - 57° W, 33°40' - 34°50' S (see Fig. 1). The natural boundaries are low hills ("*cuchillas*") of height below 300 m. Grass is the main coverage and most of the trees lie in the margins of rivers. The most important one is the Santa Lucía river, with 225 km length, being its main contributions from the San José river, of 111 km length, and the Santa Lucía Chico river, of 122 km.

The region has no dry season, and can be classified as Koeppen´s Cfa category. The average mean temperature is 17°C. The extreme values on annual basis are 43°C and -8°C. Total annual precipitation are 1000 mm. There are a somewhat large interannual variability, from 1600 mm (1959) to 500 mm (1916). The difference between the extreme monthly average is 100 mm (march) and 75 mm (july) which shows a regular pattern of precipitation during the year. Relative moisture ranges around 70%, being the maximum 78 % (june) and the minimum 60% (january).

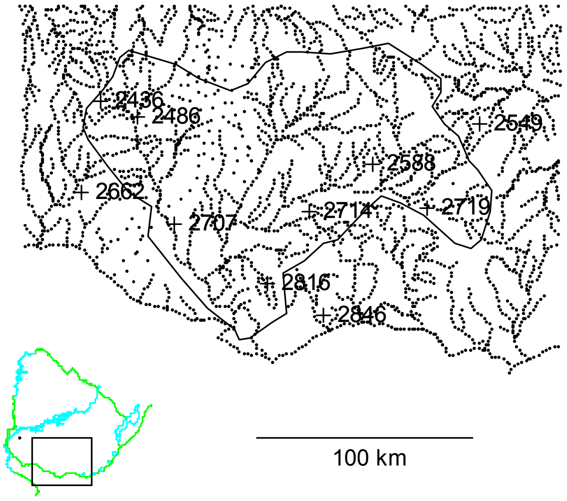

Fig. 1 Location of the pluviometric stations considered

The measurement network consists of 50 stations, mostly operated by non-technical staff from the police and the National Railway Company. The spatial arrangement suffer from such fact, because the stations have been located along railways lines, or in small villages. With a reorganization of the National Railway Company in the mid eighties, many of the railway stations in the area were shutdown, resulting in many missing records since them. We have selected 10 pluviometric stations out of 50, for the abovementioned period. Their name, code and location are given on table 1, and its layout in Fig. 1. The data taken by the operator go in two different ways to the National Weather Service. The data is collected at 7:00 AM and paper records are prepared in-situ. They are submitted by surface mail once a month. For other purposes, the information is collected by radio or telephone (where available) by the Police Department of the province, and transmitted later in the same day to the National Weather Service main office at Montevideo. Magnetic records are taken from this daily report, in a process which started some years ago. No routine check is made against the original records, even though they have been regarded as the "true" ones. The dataset have been collected, from the instrument to the paper, with almost the same routine for the whole period. The first computer appears in 1977 and records back to the beginning of the century were digitized in a short period. The task was carried out following a station order rather than a time order. So mistakes for the same day in different stations are less prone to occur.

N°	*Name*	*Latitude*	*Longitude*	*Height ASL*
852436	Puntas de Sauce	33°50'S	57°01'W	120 mts
852486	Pintos	33°54'S	56°50'W	100 mts
852549	Barriga Negra	33°56'S	55°07'W	95 mts
852588	Casupá	34°06'S	55°39'W	124 mts
852662	Cufré	34°13'S	57°07'W	92 mts
852707	Raigón	34°21'S	56°39'W	37 mts
852714	San Ramón	34°18'S	55°58'W	70 mts
852719	Ortiz	34°17'S	55°23'W	115 mts
852816	Joanicó	34°36'S	56°11'W	35 mts
852846	Olmos	34°44'S	55°54'W	40 mts

Table 1 General information of the pluviometric stations considered for the period 1960 - 1990

3. The tested mechanisms

Any mechanism to be considered might fulfill some requirements. For instance, we assume that outliers are not out of the observable range. Our situation is that we have a previously recorded database, which must have passed successfully the most trivial requirements for consistency (no negative records, etc.). So we are trying to model errors like those still remain in the dataset.

Another trivial requirement is that an error should be different from its true value. We define here the true value as the one which is hand written by the operator. We disregards other errors source, like malfunctioning of the instrument, threshold, etc. because we cannot going so far into the data.

Those before are strong requirements. We can state also weak requirements. For example we will prefer to use non parametric procedures based on available data in order to simulate outliers.

However, the error generation mechanism is strongly connected with the routines used for collection. Notice also that in many cases such routines might have changed over time. As an example, errors in the card punch process (see Coale *et al*., 1962) might not occur while typing in terminals Also automatic collection equipment (no typing!) are not error free as well, because they suffer from other mechanisms not considered here.

So this paper might be of use for datasets collected with well known, well established procedures, which might have undertaken the task of typing the paper records mostly at once and afterwards they continue keying with the same procedures (even the same software). That´s the situation in many of the third world meteorological offices. To make things worse, they collect data to be used by themselves in a) aggregated form or b) as extreme values, which are statistics very robust to non gross errors, which might stay unnoticed for long time within the dataset. Gross errors are not prone to occur because they might affect the extreme value report. No sophisticated data assimilation procedure has been available in the past, and even powerful computers are still a dream. Such National centers provide useful data which is analyzed at the global level by well skilled personal and well proven algorithms, but at the local scale most records remain as keyed, after very crude range controls.

With strong and weak requirements in mind, we can devise some different simple algorithms.

a) for random date and station, select a second random date and station until its record is different from the destination one.

b) for random date and station, select a random date for the same station until its record is different from the destination one.

We applied both mechanisms and found that the results are not satisfactory. We should notice that in the available records most (around 80.5 per cent; see table 2 for this and other summary statistics) of the values are zero. So the method will select with 80.5 per cent probability as a destination a zero rain value, and will assign there a non-zero one. This is not observed in practice, because the true zero values appear at a somewhat lower rate (24.3 per cent). We will denote as the *correct value* the one written on paper, and as *wrong value* the one available in magnetic form. So b definition there should be the same number of wrong and correct values. In case where either the wrong or the correct value is a missing one we did not include them into consideration.

In 99.8 per cent of the cases the wrong value is greater or equal than the minimum daily value. In 91.6 per cent of the cases, the wrong value is bounded by the maximum of the observations of the day, day before and day after for all the available stations. This fact was due to one-day time shift error, a somewhat typical situation due to unscheduled changes in the observation time from early morning to night. It has also been noticed that the distribution of the wrong values has longer tails than the station probability density function (pdf). So we discard the possibility of choice at random one of the neighbor values (which automatically will be between both extreme cases) and turn to choose them from a distribution of "extreme" values. Such extreme value set was built using some heuristics which will be presented later.

The days where to put the error are also not at random. The zero rain as a correct value has a substantially lower probability (24.3 per cent) than the overall population. We suggest that an error is more likely to occur in a "rainy" day than in a "dry" day, because the operator might be more aware of the typical values. We found that the probability of having less than 2.5 mm/day average rain is 81 per cent in the population, but when errors arise, such probability is around 24.0 per cent. On the other side, heavy rain events are unlikely to occur: the average is over 25 mm/day less than 2.8 per cent of the events, but when exist errors, such percentage grows up to 29.3 per cent, so we devised different mechanisms provided the daily average is below 2.5, within 2.5 and 25 and over 25 mm/day.

Summing up our model will consider some error mechanisms supported by empirical rules:

a) the wrong value is larger than the day minimum and typically is less than the observed maximum for the day, day before and day after in the catchment area.
b) we distinguish three situations "dry day", "heavy rain day" and the others, and use different sets for selecting the wrong value as well as the correct one

4. The heuristics

Despite the availability of the abovementioned rules there is still the need to specify either a pdf (probability density function) or sets where to choose randomly the correct value (to be substituted) and the wrong value (to be used instead). We add other rules in order to make the model operational, while keeping its outputs close to the observed ones. The most important rules are those regarding where to put the errors, and how many zeros to change. We divided the population in three cases according to its mean rain: (A) between 0 and 10; (B) from 10 up to 25 and (C) over 25 mm/day. From the sampled errors and for each case, we calculated the probability that the errors appear in that case, and how many of them are zero. With only such information we were able to select a day as candidate to hold an error, and to suggest within the day a specified number of non-zero readings in order to put a zero, and conversely, a number of zero readings to be substituted from values taken from other distribution.

Notice that, despite one typical mechanism for error generation is to mix the values from the day with those of the day before, we were unable to reproduce observed patterns considering directly such fact, because the errors were smaller than observed. Thus we decided to use errors obtained from a set created as follows. It contains all the minimum readings for each event, as well as all the maximum values for the current day, the day before and the day after considered for the overall population. This population has more low rain values than observed, so at random we reduced to one fourth the readings belonging to the first quartile of the non-zero values, and adjust the proportion of the zero values in order to follow the observed one. Let's denote as error set the result of this operation, which is created using information of the available dataset.

When assigning an outlier, we disregard the information of the maxima from the day before, current day and day after, and simply pick a value from the error set. What it is strictly forced is that the error should be larger than the minimum of the day (as observed in over 99 per cent of the errors).

5. Results

In table 2 some comparative results between observed/simulated are presented. It is clear from them that a perfect simulation has not been achieved. The most significant deviations were with the exactly zero rain readings. For example, on heavy rain events the model selects zero as correct value only 3.5 per cent of the conditional sample, while it has been observed 6.4 per cent. The model assigns a wrong zero value only in 29.3 per cent of the cases, while the observed conditional probability is 75 per cent. In addition to this preliminary numbers, the failure of the error model to mimic observed error behavior are also supported by the negative results of the Wald-Wolfowitz test (Gibbons and Chakraborti 1992) which analyzes two sets of numbers for the null hypothesis of belonging to the same population. Despite such pessimistic result we show the QQ-plot on figure 2 of the observed vs. simulated differences between correct and wrong values. The QQ-plot should render a linear relationship if both populations follow the same distribution but maybe with different parameters. The results are remarkably linear, except for the large differences. Figure 3 shows a comparison between the population of wrong values. The deviation is again more evident in the larger rain events, despite the cases are rather few. Figure 2 compares the population of the correct values (i.e. those values which will be replaced), and as before, the discrepancies arise with the larger rain values. So the model failed mostly with the heavy rain events, which are underestimated.

Probability of events with	Population	With-errors events
little rain (average <10)	90.8	29.1/30.5
heavy rain (average >25)	2.8	29.3/28

Probability of readings with:		Correct values	Wrong values
exactly zero rain	80.5	24.3/25.1	68.8/51.8
zero rain in little rain events (avg. < 10)	86.9	58.3/59.9	42.4/59.8
zero rain in heavy rain events (avg. > 25)	7.0	6.4/3.5	75.0/29.3
rain over the day minimum			99.8/99.6
rain below three day's maximum			91.6/80.5

Table 2 Some statistics of the simulated and observed population. The X/Y denotes observed X and simulated Y values.

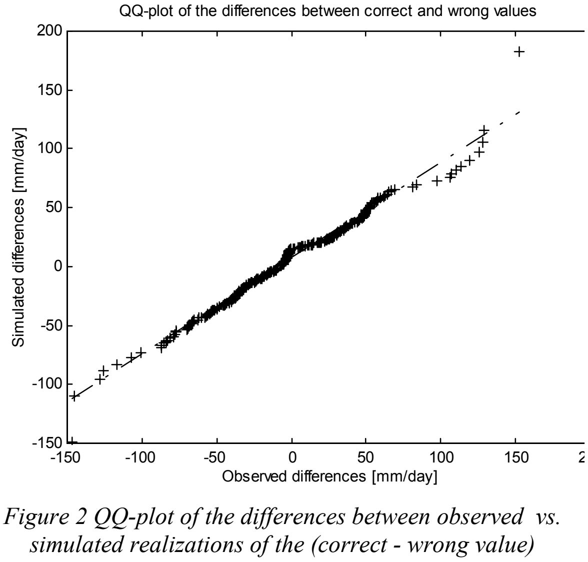

Figure 2 QQ-plot of the differences between observed vs. simulated realizations of the (correct - wrong value) population. The simulated population has 6792 events, while the observed one has 478.

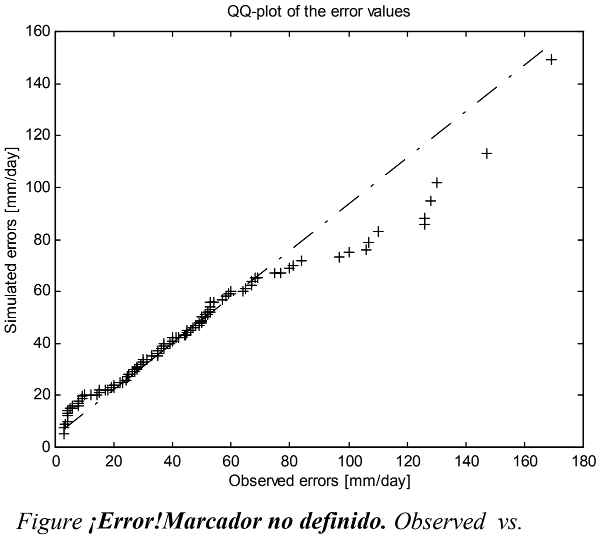

*Figure **¡Error!Marcador no definido.** Observed vs. simulated QQ-plot of the wrong values. The simulated population has 6792 events, while the observed one has 478.*

For the sake of completeness, we have also make some runs with a direct implementation of MIXCAR (MIXed Completely At Random) which in practice requires to choose the wrong value from any other station and date; the only strong requirement was that the wrong value should be different from the correct one. This procedure has the nice property that the distribution of the overall population is barely affected. The results were comparable with those presented here in terms of the correct value (i.e. where to put the outlier) but they were deceptive in terms of the difference between wrong and correct, and also the distribution of the wrong values, which were very different from the quasi-linear appearance of the abovementioned figures. The alternative of using values from the same station selected at random were also tested, and it performed even worse.

6. Discussion and conclusions

The problem of generate random replications of a daily rain database with similar outliers observed in practice has been addressed. Using empirical facts regarding some of the typical outlier-generation mechanism we described both the schema and its results. The model underpredicts the occurrence of heavy rain readings, which are not so common but occurs in practice, which led in turn to fail in passing some statistical tests. However, from a visual comparison of the QQ-plots which should be linear for perfect fit, the results are clearly of use, and no major effort towards improvement were done. The procedure classifies the events in three categories A, B and C ranging from light to heavy daily mean rain, and forces to honor the proportion of observed outlier occurrence for each category, as well as the proportion of errors appearing in zero rain correct values. The wrong values are chosen at random from a set created using the available information, which allow to apply this procedure to different databases. Such set is tailored in order to produce results as similar as possible with the ones observed. Two more alternatives have also been evaluated: MIXCAR(MIXed Completely At Random) and simply mix the records for the station. Despite that both generate errors without modifying the overall pdf of the population, they failed to model properly the wrong value and the difference with the observed one, so they should be discarded. Since the observed errors have appear after the application of error detection procedures (López *et al.* 1994) it has been investigated to which extent the errors found were an artifact of the procedure itself. This has obvious implications, because the error generation code will be used to test the ability of the error detection procedure. So one year were carefully typed once more, and all discrepancies were corrected; this is denoted as Duplicate Performance Method (Minton 1969) and is assumed that it will highlight both gross and subtle errors as well. The conclusions show that few errors still exist for that year after the previous depuration process, so we were confident about the representativeness of the observed errors.

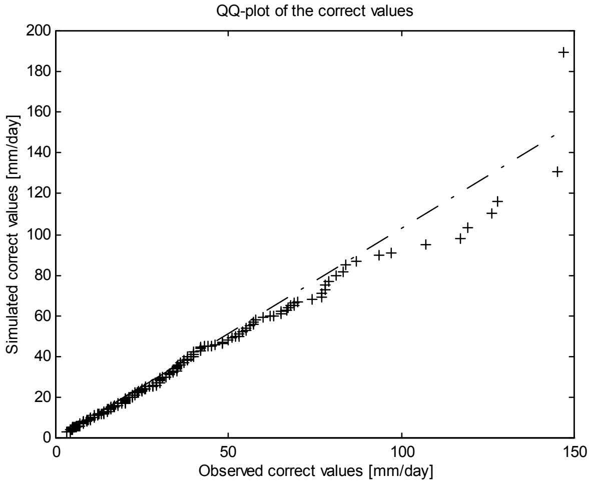

Figure 2 Observed vs. simulated QQ-plot of the wrong values. The simulated population has 6792 events, while the observed one has 478.

References

Di Mego, 1988, “The National Meteorological Center Regional Analysis System” *Mon Wea Rev* **116**, 977-1000

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 data" In *Proceedings of A workshop on ERS-1 wind and wave calibration*, Schliersee, FRG, 2-6 June, 1986 (ESA SP-262)

Gandin, L. S., 1988, “Complex Quality Control of Meteorological Observations” *Mon Wea Rev*, **116**, 1137-1156

Gibbons, J. D. and Chakraborti, S., 1992, “Nonparametric Statistical Inference” 3rd. Edition. Marcel Dekker Inc. ISBN 0824786610.

Hawkins, D. M., 1993, “The feasible set algorithm for least median of squares regression” *Computational Statistics & Data Analysis*, **16,** 81-101.

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" *Mon Wea Rev*, **114**, 5, 861-879.

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

Minton, G., 1969, “Inspection and correction error in data processing” *JASA*, **64**, 328, 1256-1275

Parrish, D.F. and Derber, J.C., 1992, "The National Meteorological Center`s Spectral Statistical Interpolation Analysis System". *Mon Wea Rev*, **120**, 1747-1763.

Reek, T.; Doty, S. R. and Owen, T. W., 1992. “A Deterministic Approach to the Validation of Historical Daily Temperature and Precipitation Data from the Cooperative Network” *Bull. Amer. Met. Soc.*, **73**, 6, 753-762

Rocke, D. M. and Woodruff, D. L., 1996, “Identification of outliers in Multivariate Data” *JASA*, **91**, 435, 1047-1061