PUBLS. INST. GEOPHYS. POL. ACAD. SC., C-99 (398), 2007

The Hourly Mean Computation Problem Revisited

Jean-Jacques SCHOTT [1] and Hans Joachim LINTHE [2]

[1] Ecole et Observatoire des Sciences de la Terre
5 rue Descartes, F-67084 Strasbourg Cedex France
e-mail: JeanJacques.Schott@eost.u-strasbg.fr

[2] GeoForschungsZentrum – Adolf Schmidt Observatory
Lindenstrasse 7, D-14823 Niemegk Germany
e-mail: linthe@gfz-potsdam.de

Abstract

The computation of hourly means based upon one minute values raises the question of the influence of gaps on the mean values. The issue is revisited in this paper from a statistical point of view. In a first step, relevant statistics for gap and data distribution within hourly intervals are built up using actual time series. Then, the statistics are applied to full one minute field values in order to evaluate the dispersion of the hourly means due to the gap and data statistical distribution. The data from PAF observatory (Port-Aux-Français, Kerguelen Island) in the range 1999 to 2003 is used for illustration.

1. Introduction

Hourly means have been used for a long time in external field modeling, for instance diurnal variation modeling. In recent years, they have become increasingly interesting in internal field or internal plus external field modeling (i.e., Sabaka *et al*. 2002). However, a long standing problem has come up with the hourly mean computation, namely the tolerable length of gap within an hourly interval. This issue has become particularly acute since the advent of digital recording. Few references on this problem are available. The most recent paper was published by Mandea (2002). It was motivated by a question addressed by L. Loubser (Hermanus observatory) to the community of observers. Among the answers, the one given by De Santis could be considered a starting point for the present study: “In my opinion, for most typical kinds of quiet and moderate magnetic activity, it would be probably better to have one data every three (i.e. 33% data only) instead of having just the first half of the whole period but nothing of the remaining (i.e. 50 % data).” This opinion strongly suggests a statistical approach of the issue.

2. Elaboration of Hourly Mean Statistics

An overview of the data and gaps distribution is readily provided by the scrutiny of one minute time series over a selected time span. Figure 1, left, shows the histograms built up using the X, Y, Z (in geographical reference frame) and F components from PAF observatory over the years 1999-2003. In order to obtain representative distributions, we should select time intervals free of major changes in the instrumentation and/or data processing methods, etc. The distribution of data segments is equivalently the distribution of time elapsed between two successive failures. This strongly suggests a Weibull distribution (Gibra 1973) designed to model electronic component life time, and therefore relevant to model the failures arising in data acquisition equipments. On the other hand, the distribution of gaps is close to a mere exponential distribution. This depends on a variety of parameters as a result of the way the observatory is operated.

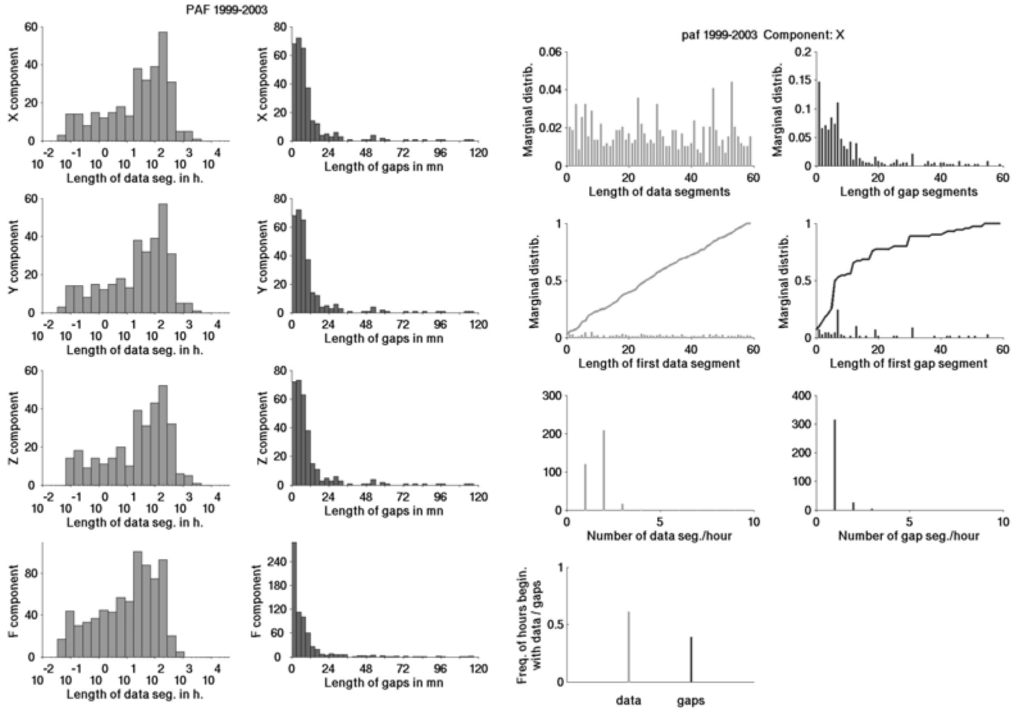

Fig. 1. Left: distribution of the length of data segments (left) and gap segments (right) for PAF observatory during the years 1999-2003. The horizontal scale is logarithmic for the data and linear for the gaps. Right: statistics regarding the distribution of data segments and gap segments within hourly intervals. See text for explanations.

Regarding hourly mean computation, the relevant basic information is the hourly intervals containing both data and gaps. A large number of situations may be imagined. Elementary combinatorial analysis demonstrates that there are about 1.15×10^{18} possibilities. However, all of them are not realistic. It is more advisable to base the statistics upon real distributions. In any case, the data/gap, denoted T_{dg} in the sequel, or gap/data transitions, denoted T_{gd}, are the considered random variables.

Therefore, the first step is, in a given series of minute values, to select the hourly intervals containing both data and gaps. Indeed, for obvious reasons, intervals without gaps or data can be discarded. Preliminary statistics may be performed on the distribution of T_{dg} and T_{gd}. Figure 1, right, shows some distributions of interest. Upper left: distribution of length of data segments; upper right: distribution of length of gap segments; second row left: length of the first data segment in the hourly interval (histogram and cumulative function); second row right: length of the first gap segment (histogram and cumulative function); third row left: number of data segments per hour; third row right: number of gap segments per hour; bottom: frequency of segments beginning with data (light gray), respectively gaps (dark gray). In the example of PAF observatory, the length of data segment is roughly uniform, whereas the length of gap segments clearly is not. The third row illustrates the fact that not all theoretical distributions are realistic: for instance, the number of failures per hour hardly overruns three.

Having obtained a set of hourly intervals, whose distribution of data and gaps is typical for a given observatory and period of time, the next step consists in computing the estimated hourly means based upon this set, on one hand, on full records of one minute values on the other hand. The estimated hourly means are random variables whose distribution depends on the maximum tolerated gap length and the shape of the field variation, as already noticed by Mandea (2002). The shape itself depends on the magnetic activity and the location of the observatory, mainly its geomagnetic latitude. The hourly means being regarded as random variables, their distribution may be represented on histograms and characterized by usual statistical parameters like their standard deviation, percentiles (less sensitive to outliers).

3. Examples

3.1 Hourly mean distribution for a quiet day

Figure 2 shows an example of the hourly mean distribution obtained for PAF observatory based upon the distribution of data and gaps segments computed using the method outlined in the previous section, on one hand, and the daily records for 3 February 2001, on the other hand. Left part of Fig. 2 displays the distribution of the hourly mean for hour 7 U.T. (the hours are numbered from 0 to 23) and shows the increase of dispersion due to the increase of gap length. The lower drawing displays the daily variation of the selected component, in this case the component along geographic east. The field variation during hour 7 is thickened.

The head “at most *n* minutes missing” means that all distributions of data and gap sequences with less than n+1 missing values are taken into account. Therefore, the number of suitable cases increases with the tolerated gap length. std stands for standard deviation and the 95% figure is the width of the interval containing 95% hourly means. Under a Gaussian assumption, it would be twice the standard deviation. It can be noticed that this is roughly the case, although a Gaussian assumption is not required in this study. Half of the 95% interval may also be interpreted as a confidence interval at the 95% confidence level. Right part of Fig. 2 is like the left part apart from the maximum tolerated length of gap being fixed and its influence checked across the day using 6 selected hours highlighted on the lower drawing. Thus, the figure illustrates the influence of the shape of the field variation. Although the magnetic activity is very small (e.g. K indexes on Fig. 3), a moderate variation of the hourly mean scattering may be noticed. It is clearly correlated with the slope of the hourly variation. Hour 7 selected previously has the largest slope and, hence, the largest dispersion.

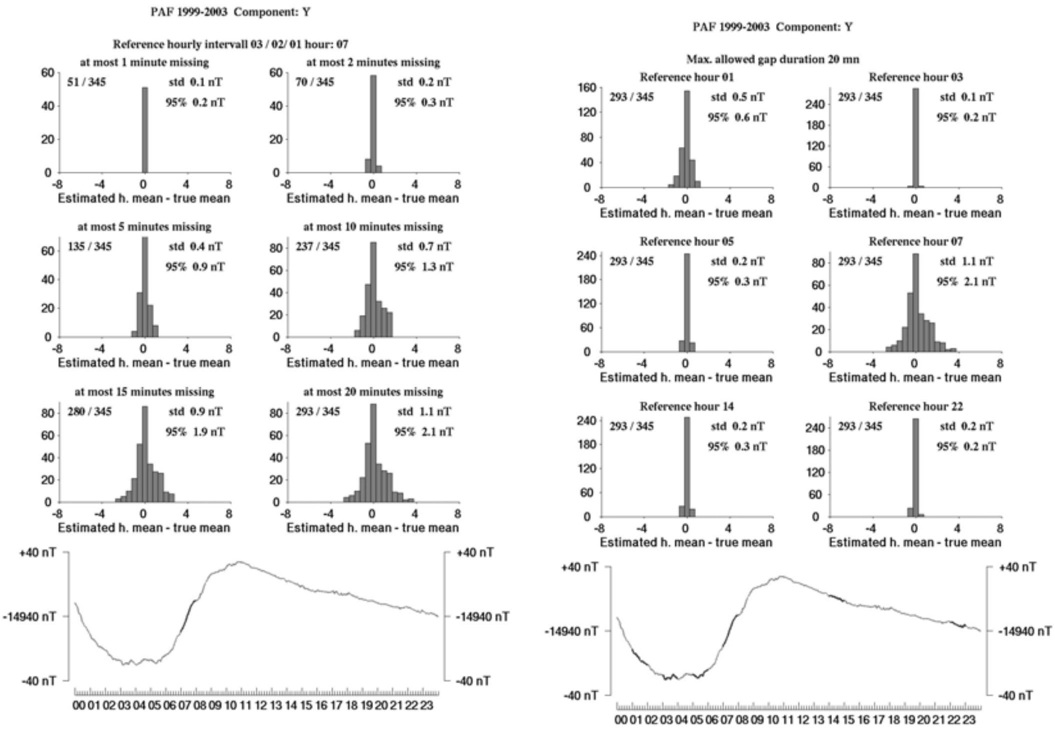

Fig. 2. Left: distribution of estimated hourly means around the true mean for various maximum gap lengths, the hourly interval being fixed. Right: similar to left with a fixed maximum length of gap and various hourly intervals throughout the day. See text for further explanations.

The head “at most n minutes missing” means that all distributions of data and gap sequences with less than n+1 missing values are taken into account. Therefore, the number of suitable cases increases with the tolerated gap length. std stands for standard deviation and the 95% figure is the width of the interval containing 95% hourly means. Under a Gaussian assumption, it would be twice the standard deviation. It can be noticed that this is roughly the case, although a Gaussian assumption is not required in this study. Half of the 95% interval may also be interpreted as a confidence interval at the 95% confidence level. Right part of Fig. 2 is like the left part apart from the maximum tolerated length of gap being fixed and its influence checked across the day using 6 selected hours highlighted on the lower drawing. Thus, the figure illustrates

the influence of the shape of the field variation. Although the magnetic activity is very small (e.g. K indexes on Fig. 3), a moderate variation of the hourly mean scattering may be noticed. It is clearly correlated with the slope of the hourly variation. Hour 7 selected previously has the largest slope and, hence, the largest dispersion.

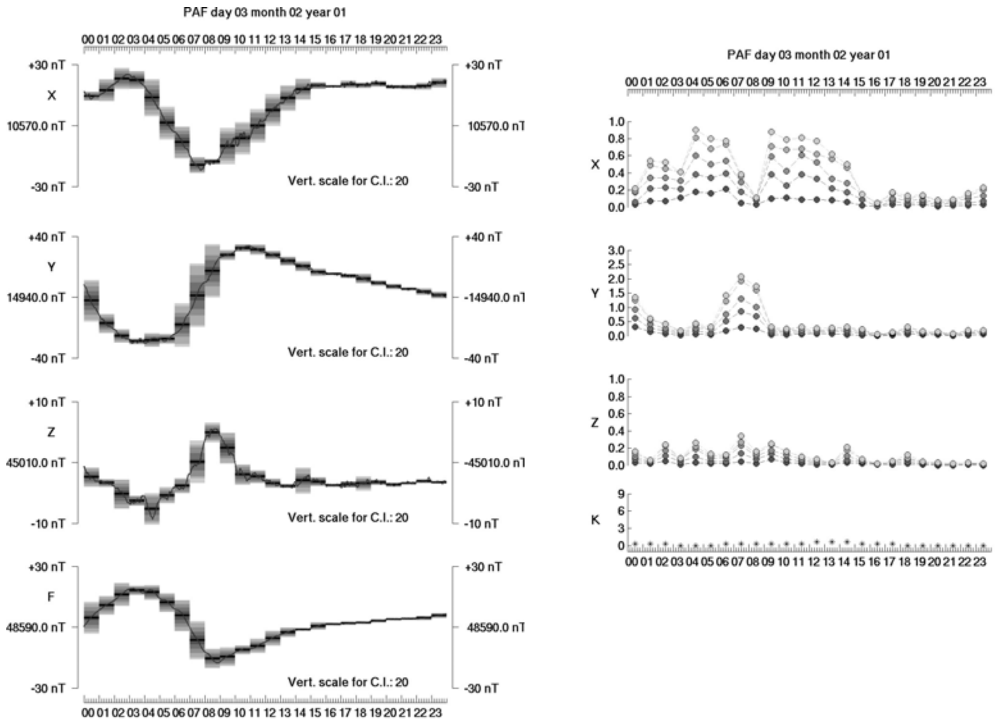

Fig. 3. Left: Confidence intervals for the 24 hourly means and tolerated gap lengths 2, 5, 10, 15, 20 minutes. Colours vary from dark grey to light grey accordingly. Vertical scale of the c.i. magnified 20 times. Right: width of the confidence interval versus hour of the day, for the same maximum tolerated gaps as on the left figure. Bottom: K indexes for the day.

Figure 3, left, displays the confidence intervals for the 24 hourly means and tolerated gap lengths 2, 5, 10, 15, 20 minutes. The confidence intervals are pictured as rectangles colored from dark grey (2 minute gaps) to light grey (20 minute gaps). For the sake of clarity, the confidence intervals are magnified 20 times with respect to the vertical scale. Again, we see the clear correlation between the slope of the hourly variation and the width of the confidence interval. Figure 3, right, summarizes the confidence interval widths (note that the vertical scales are variable), for every hour and the same five selected gap lengths as on the left. The lower drawing shows the variation of the K indexes over the day. There is no correlation between the width of the confidence intervals and K, all the lesser as the K index is a logarithmic scale not designed for characterizing quiet day variations. Once again, the diurnal variation of the confidence interval is correlated with the slope of the field diurnal variation, which could be modeled with a Sq model (Malin and Gupta 1977, for instance).

3.2 Hourly mean distribution for a disturbed day

On the other end of the field disturbance scale, Figs. 4-5 illustrate the effect of strong disturbances on the computation of hourly means. The selected day is within the so-called Halloween storm. Figure 4 makes evident the dramatic increase of the hourly mean scattering. Note the large variation of the horizontal scale, either according to the tolerated gap length or to the shape of the hourly variation. In addition, the figure reveals the poor sampling of these very scattered distributions, despite the long time series analysed. Figure 5, left, is similar to Fig. 3, left, except for the vertical magnification of the confidence intervals, reduced to 10 times the field vertical scale. It shows also that even during strongly disturbed days, some hourly means are moderately scattered. Finally, Fig. 5, right, equivalent to Fig. 3, right, summarizes the daily variation of the confidence intervals and shows again that the correlation with the K indexes is rather poor in this case. The lack of correlation may be explained by the fact that during strong magnetic storms, PAF behaves like an auroral observatory due to the northwards expansion of the auroral oval. This fact is supported by the clear correlation observed, on the other hand, with the auroral indexes (Fig. 6).

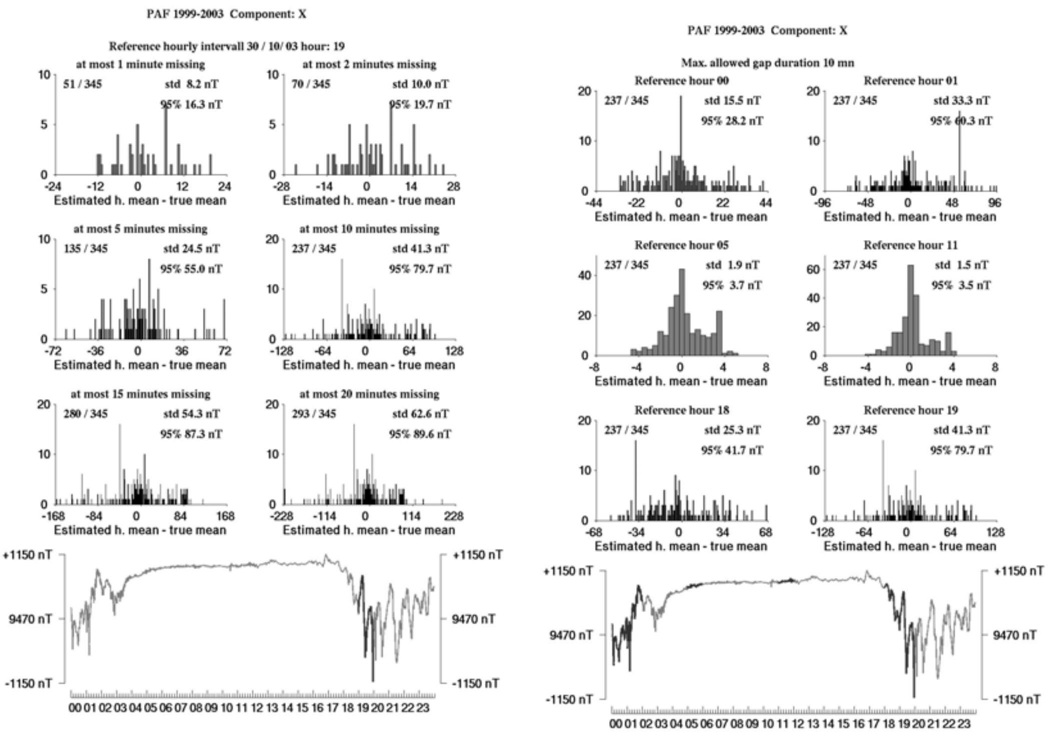

Fig. 4. Same as Fig. 2 for a disturbed day, apart from the test hour selection.

4. Discussion and Further Developments

One motivation for this study was the hope that a simple policy for computing the hourly means or not, according to the number of data missing, could be suggested.

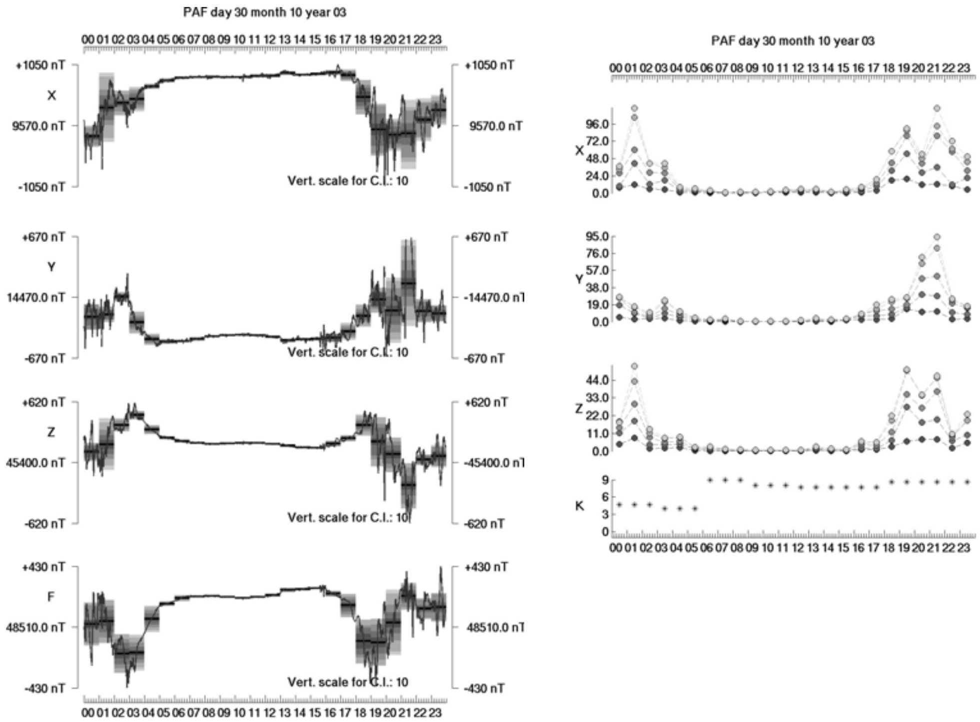

Fig. 5. Similar to Fig. 3 for the disturbed day used as an example in Fig. 4. On the left, vertical scale for the confidence intervals magnified 10 times.

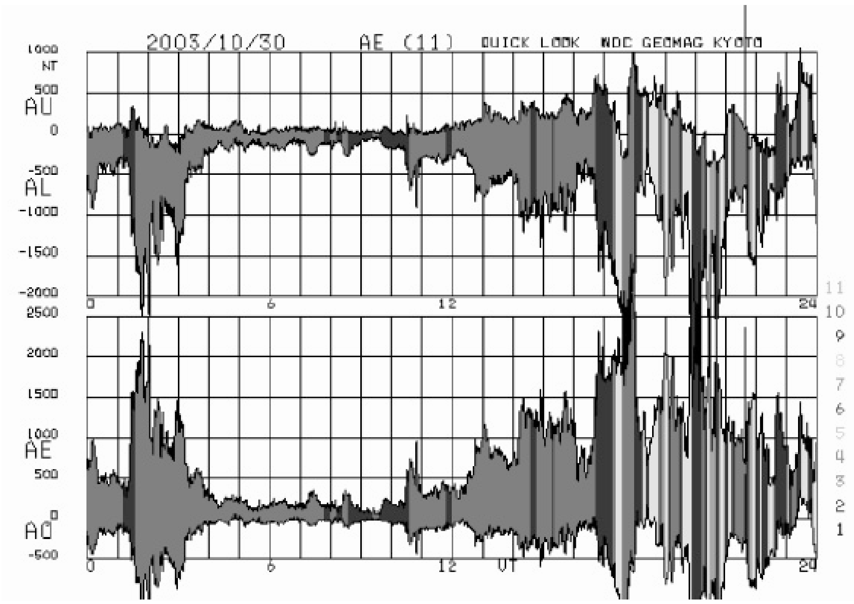

Fig. 6. Quick look auroral indexes for the 30 October 2003 (downloaded from web site swdcwww.kigi.kyoto-u.ac.jp). To be compared to the daily variation of the confidence intervals in Fig. 5.

This goal clearly could not be reached, due to the large number of parameters involved. These parameters include the statistical distribution of data and gaps, which is peculiar to each observatory, the location of the observatory, the level of magnetic disturbance. The comparison of Figs. 5 and 6 suggests that appropriate magnetic indexes could be a relevant parameter for taking into account the magnetic activity, above some level of disturbance. On the other end of the disturbance scale, Sq models might be helpful to play an equivalent role. Anyway, assuming that the distribution of data and gaps, and of course the location of a specific observatory, would be known, the upper threshold for missing data, with respect to an upper limit of error assigned to the hourly means, should be a function of an hourly parameter reflecting the shape of the field variation. In addition, this parameter would depend on the component, that is would be a 3 or 4 dimension vector.

On the other hand, this study draws the attention on the fact that the hourly mean based upon an incomplete set of data is only an estimate of the true, unknown value. It can therefore be treated as a random variable, whose properties may be investigated with the method outlined above. In particular, a confidence interval may be associated to each hourly mean. In order to compute it, either a typical record (without gaps) or an appropriate parameter, fairly accurately correlated with the confidence interval in order to provide a reasonable estimation for it, has to be available. To this end, the analysis of a fairly large number of suitably distributed observatories and levels of disturbance would have to be performed.

Yet, the statistical analysis presented herein is rather elementary from at least two points of view. First, as can be seen in Fig. 4, the distribution of hourly means, based upon actual sequences of data and gaps, is poorly sampled, due to the rather small size of the set of hours containing gaps. This is probably (and hopefully) the case for most observatories. This observation calls for the search of a method which could generate hourly sequences of gaps and data having the same statistical properties than the actual ones. We have attempted to do it using a Markov chain process based upon transition probability matrices data/gap and gap/data constructed with real data. The result is not really satisfying because the transition probability matrices themselves are poorly sampled. In an attempt to build up smooth versions of them, we have adjusted the empirical transition probability functions by theoretical distributions, namely a mixture of positive binomial and negative binomial distributions (Gibra 1973). The result is better, but not significantly. Thus, further investigations have to be carried out, although we believe that they should not change significantly the confidence interval estimates.

A completely different way to explore would be to implement statistical methods dealing with missing data (Little and Rubin 1987). The idea behind this concept is to find a way of replacing the missing data without altering the estimates of the parameters looked for – in the present case, merely the mean value. For instance, the field variation in every incomplete hourly interval could be modeled by some parametric function (polynomials, sine functions, etc.) whose parameters would be estimated from the available data by some least-squares or likelihood method, according to schemes referred to in the statistical literature (see for instance Little and Rubin 1987, for references). This way of computing hourly means would perhaps not completely make meaningless the question of the tolerable length of gap, but would certainly make less critical the choice of the threshold.

5. Conclusion

The influence of missing data on the hourly means of the magnetic field components has been investigated from a statistical point of view based upon actual distributions of gaps and data peculiar to a specific observatory. The study confirms and amplifies conclusions already drawn by Mandea (2002), namely that the error involved depends, beside the length of gap, on the shape of the field variations and the level of the magnetic disturbance. These features, in turn, depend on the position of the observatory. The influence of the magnetic disturbance upon the confidence interval might be quantified with the help of an appropriate magnetic activity index. However, overall, the problem of fixing a limit to the tolerable length of gap in an hourly set of data is probably ill-posed. One way of circumventing the difficulty would be to rely on statistical methods dealing properly with missing data.

References

Gibra, I.N., 1973, *Probability and Statistical Inference for Scientists and Engineers*, Prentice-Hall, London, 596 pp.

Little, R.J.A., and D.B. Rubin, 1987, *Statistical Analysis with Missing Data*, John Wiley & Sons, New York, 278 pp.

Malin, S.R.C., and J.C. Gupta, 1977, *The Sq current system during the International Geophysical Year*, Geophys. J.R. astr. Soc. **49**, 515-529.

Mandea, M., 2002, *60, 59, 58, … How many minutes for a reliable hourly mean?*, Proceedings Xth IAGA Workshop, Hermanus 112-120.

Sabaka, T.J., N. Olsen and R.A. Langel, 2002, *A comprehensive model of the quiet-time, near-Earth magnetic field: phase 3*, Geophys. J. Int. **151**, 32-68.

Accepted February 14, 2007