DNMI DET NORSKE METEOROLOGISKE INSTITUTT

klima

THEORETICAL ANALYSIS OF THE DIP-TEST IN QUALITY CONTROL OF GEOPHYSICAL OBSERVATIONS

PETTER ØGLAND

RAPPORT NR. 24/93 KLIMA

DNMI - RAPPORT

DET NORSKE METEOROLOGISKE INSTITUTT
POSTBOKS 43 BLINDERN 0313 OSLO 3

TELEFON: 22 96 30 00

ISBN

RAPPORT NR.

24/93 KLIMA

DATO
14.06.1993

TITTEL

THEORETICAL ANALYSIS OF THE DIP-TEST IN QUALITY CONTROL OF GEOPHYSICAL OBSERVATIONS

UTARBEIDET AV

Petter Øgland

OPPDRAGSGIVER

DNMI

SAMMENDRAG

A simple mathematical test for recognizing certain types of errors called dips i series of samples from continuous geophysical parameters is described.

The test is generalized in respect of handling time series with missing values.

The test is being further analysed and found to be a special case in a class of dip-tests. The test is however not optimal, but can be replaced with a test which is equally simple.

Some statistics are provided for comparing the improved dip-test with the original.

UNDERSKRIFT

Petter Øgland	Bjørn Aune
FORSKER	FAGSJEF

THEORETICAL ANALYSIS OF THE DIP-TEST IN QUALITY CONTROL OF GEOPHYSICAL OBSERVATIONS

1. INTRODUCTION

While performing statistical analysis on hourly observations of climatic elements for weather station 8711 Andøya [1], there was a need to develop some algorithms controlling the quality of the data. These algorithms had to be easy to implement and reasonably efficient in detecting suspicious geophysical values. This report documents some of the problems and results concerning developing and improving an algorithm called the dip-test, used in the analysis of continuous physical parameters.

The author is also participating in a team developing a relational database system for geophysical data. One of the major ambitions of the team is to establish a control system with reasonable efficiency and security. Such a system will include different kinds of tests in different parts of the data flow, including the dip-test described in this report. A specification of the control system is provided in [2].

The author wishes to thank P. Ø. Nordli, E. J. Førland and K. A. Iden at DNMI - Klimaavdelingen for their interesting remarks.

2. DEFINITIONS AND EXAMPLES

2.1 The dip-test

The dip-test, as used in the software package designed for Aanderaa automatic stations in use at DNMI, was developed by A. Skartveit at Geofysisk Institutt Avd. B, Universitetet i Bergen.

The test is a simple one, and is used to find outlayer values in series of any geophysical parameter with a continuous distribution. The sampling ratio must be constant.

The dip-test is formally defined as follows:

Definition 2.1.1

Given a positive real number δ depending on the physical parameter $x(t)$ in question, an observation $x_i = x(t_i)$ which satisfies the condition

$$(x_{i-1} - x_i)(x_{i+1} - x_i) > \delta^2 \qquad (2.1.1)$$

is regarded as suspicious and may be rejected. ◇

The test can be interpreted as the product of the leap $x_i - x_{i-1}$, in the time interval between t_{i-1} and t_i, and the leap $x_{i+1} - x_i$ in the time interval between t_i and t_{i+1}. The dip-test acts, due to the product of the leaps, as described below:

1. If each leap is of moderate size, greater than delta, and in opposite direction, the value is regarded as suspicious.

2. If the leaps are in opposite directions, and one of the leaps is of greater size than the other in such a way that the product is greater than delta square, the value is regarded as a suspicious one.

One should however notice that the second situation also implies that no matter how great one of the leaps is, if the other is sufficiently small, the value will not be looked upon as suspicious.

Example 2.1.2

The dip-situation is illustrated by figure 2.1.1 below. In the figure a sample of hourly observations of relative humidity from weather station 8711 Andøya is displayed. A punching error has occurred as the observation of 84% is digitalized as 8%. ◇

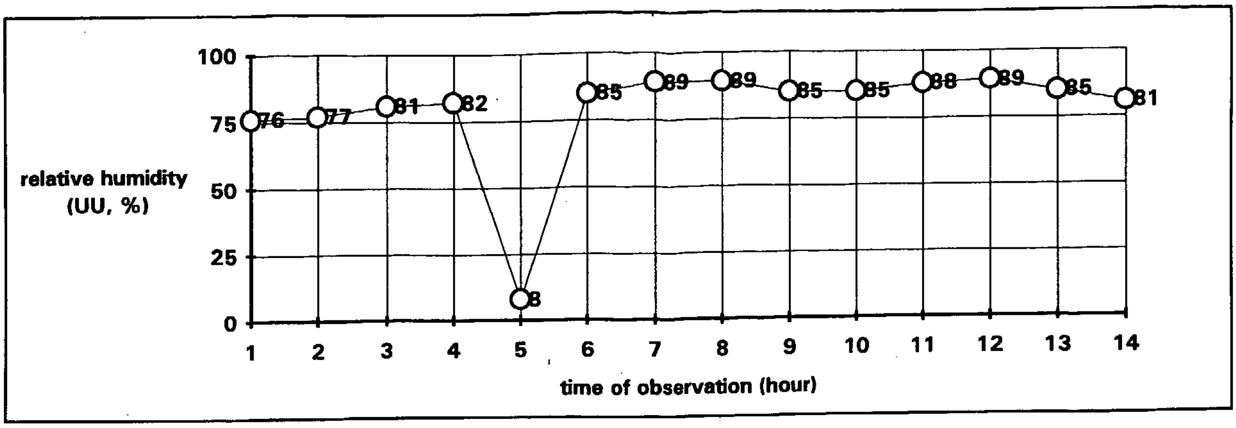

Figure 2.1.1 A dip in a sequence of hourly observations of relative humidity

In recognizing dips as the one in figure 2.1.1 as suspicious, one must have a practical knowledge of what kind of dips that are probable to occur for a given parameter at a given location, and what kind of dips that are improbable.

Example 2.1.3

As practical values at DNMI one has empirically selected values $\delta^2 = 900\ (\%)^2$ for relative humidity (UU), $\delta^2 = 55.7\ (m/s)^2$ for wind speed (FF) and $\delta^2 = 25.0\ (°C)^2$ for temperature (TT). ◇

2.2 The rejection area

By denoting $x = (x_i - x_{i-1})$ and $y = (x_i - x_{i+1})$, the dip-test (2.1.1) is associated with the implicit curve

$$xy = \delta^2, \qquad (2.2.1)$$

that is the hyperbole scaled with a factor δ^2. The curve can be used to define the rejection area associated with the test.

Definition 2.2.1

For a test T of the type where an observation x_i is rejected due to $T(x_i - x_{i-1}, x_{i+1} - x_i)$ being greater than some constant δ independent of x_i, the area

$$\Omega = \{(x,y): T(x,y) > \delta\} \qquad (2.2.2)$$

is defined as the rejection area of the test. ◇

The rejection area associated with a test plays a fundamental role in the analysis to come, especially when comparing the dip-test with other tests of a similar nature. Before this however, the concepts introduced so far will be illustrated by an example.

Example 2.2.2

Figure 2.2.1 illustrates the rejection area for relative humidity.

Given the delta-values described in example 2.1.3, observations that generate values close to the boundary of the rejection area may be correct and may be incorrect. For the parameters in question (FF, UU, TT) and delta values given in example 2.1.3, there is a rough fifty-fifty percent chance that suspicious values are in fact erroneous. ◇

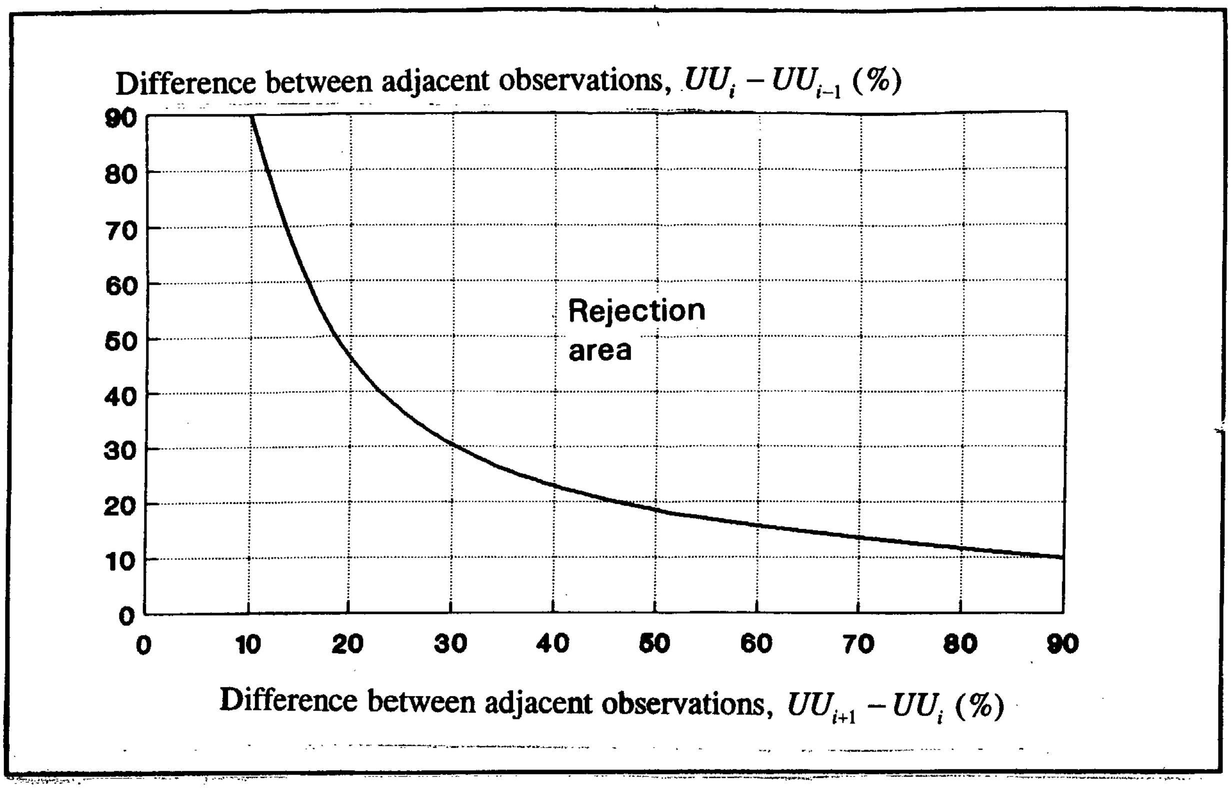

Figure 2.2.1 Rejection area for dip-test of relative humidity

3. IRREGULARLY ARRIVING OBSERVATIONS

One major disadvantage with formula (2.1.1) is that the test depends on a complete series of data, that is observations must be arriving uniformly in time and without missing values.

If the data set is sampled uniformly in time and the set contains missing values, these have to be interpolated before the test can be used. The test can however be partially applied in incomplete series wherever possible.

3.1 Non-uniform sampling

The weaknesses of dependability of uniform sampling in time can easily be overcome.

Generalization 3.1.1

Given a positive real number δ, an observation $x_i = x(t_i)$ should be marked suspicious if

$$\left(\frac{x_{i-1} - x_i}{t_i - t_{i-1}}\right)\left(\frac{x_{i+1} - x_i}{t_{i+1} - t_i}\right) > \delta^2 \qquad (3.1.1)$$

◇

It is observed that (3.1.1) is reduced to (2.1.1) if the samples are taken with uniform time intervals of magnitude one.

It should also be noticed that the formula (3.1.1) can be regarded as an approximation of the product of the right hand- and left hand-side derivative of the observation x_i in the parameter from where x_{i-1}, x_i, x_{i+1} is sampled.

3.2 Uniform sampling with missing values

The generalization (3.1.1) is of practical importance if one does not want to interpolate missing values before using the test. If however (3.1.1) is used in an implementation of the dip-test, modular arithmetic must be applied.

As an alternative approach one can use the following simplification:

Generalization 3.2.1

If samples of a continuous parameter is taken uniformly in time, that is $t_{i+k} = t_i + k$, *the formula (3.1.1) is reduced to*

$$\frac{(x_{j-m} - x_j)(x_{i+n} - x_j)}{mn} > \delta^2 \qquad (3.2.1)$$

where x_{j-k}, $k = 1,2,\ldots,m-1$ and x_{j+r}, $r = 1,2,\ldots,n-1$ *corresponds to missing values.*
◇

It should however be noticed that when applying formula (3.1.1) or (3.2.1) in a data set containing missing values, the greater the size of the gaps in the data set the less is the chance of finding errors. As a general rule, the dip-test should only be applied in series of observations with good data coverage.

4. RELATED TEST FORMULAS

As the dip-test consists of a product of two leaps it is easy to make sure that the leaps are in opposite directions. That the test is well designed, when it comes to detecting dips of different kinds, is not so sure.

The discussion below starts with an observation regarding the dip-test's relationship with the geometrical mean. The test will serve to be a leap-test in certain circumstances, which leads us to investigate a class of leap-tests in order to find out whether it is possible to combine the concepts of leaps and dips in a prolific manner.

The rejection areas play an important role in the investigation of the tests as it turns out that these areas should be convex in order to preserve some desirable properties of the original test. The original dip-test is found to be a member in a family of convex dip-tests. In this family there is a test equally simple to the original, but with a theoretically better design.

4.1 The dip-test's relationship with geometrical mean

The two situations described in the first section, concerning the efficiency of the dip-test, indicated that the test is a sort of mean between leap-situations and dip-situations. This presumption can be made more clearly with the following observation:

By denoting $x = (x_i - x_{i-1})$, $y = (x_i - x_{i+1})$, and noticing that the product xy is always non-negative, the dip-test can be described as

$$\sqrt{xy} > \delta \qquad (4.1.1)$$

In other words, if the geometrical mean (38f, [3]) is greater than the tolerance delta, the observation should be regarded as suspicious.

The significance of this observation is that the dip-test is also a sort of leap-test as the value in question is regarded as suspicious if the mean leap is greater than some tolerance.

The following section investigates the idea of leap-tests and the idea of combining the concept of a leap-test and a dip-test.

4.2 Leap-tests

The leap-tests in this exposition are symmetrical and consist of different ways of measuring the variation around the point x_i by means of sums and exponents.

Definition 4.2.1

For every number $p \geq 1$ *the formula*

$$\left(|x_{i-1} - x_i|^p + |x_{i+1} - x_i|^p\right)^{1/p} > \delta \qquad (4.2.1)$$

defines a test where the value x_i *should be regarded as suspicious if the tolerance delta is exceeded.* ◇

Figure 4.2.1 displays the curves and rejection areas generated by formula (4.2.1) in correspondance with definition 2.2.1.

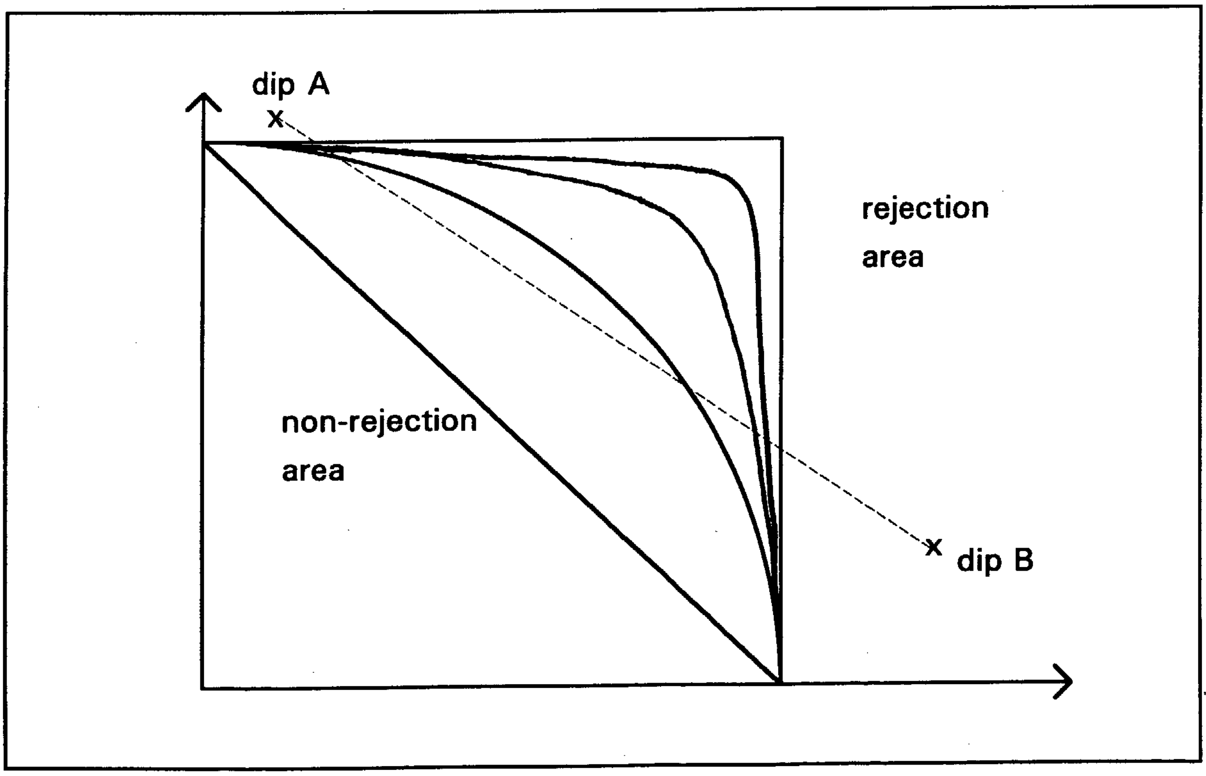

Figure 4.2.1 Rejection areas for the leap-tests.

One of the main features of the curves, apart from being symmetrical across the line $x=y$, is that for all $p > 1$ non of the curves are convex. This has the following fatal consequence:

One can find a pair of dip situations (dip A, dip B cf. fig. 4.2.1) where both dips consist of a great leap and a small leap, both subject to rejection, but such that there also exists situations, corresponding to points on the line between the two points in the rejection area (dip C $= \lambda A + (1-\lambda)B$, for a $\lambda \in (0,1)$), that would cross the non-rejection area.

The points on the line that cross the non-rejections area corresponds to dip-situations where at least one of the leaps is greater than the smallest leap in the original pair of dips.

As this situation is rather unattractive when searching for dips, the tests of the form (4.2.1) are ignored apart from the test corresponding to $p=1$.

Proposition 4.2.2

It is not possible to combine the concepts of a dip-test and a leap-test as defined by (4.2.1) with $p > 1$ if the dip-test is to be consistent.

◇

The consequence of proposition 4.2.2 is that the mixing of the concepts of dip-tests and symmetrical leap-tests is not prolific, and consequently a study of tests where the dip is more in focus will be ensued.

4.3 The class of convex dip-tests

As the convexity is a desired property when searching for dips, the class of symmetrical convex curves associated with such tests should be investigated.

In figure 4.3.1 there is a sketch made of curves that inhabit the property of convexity.

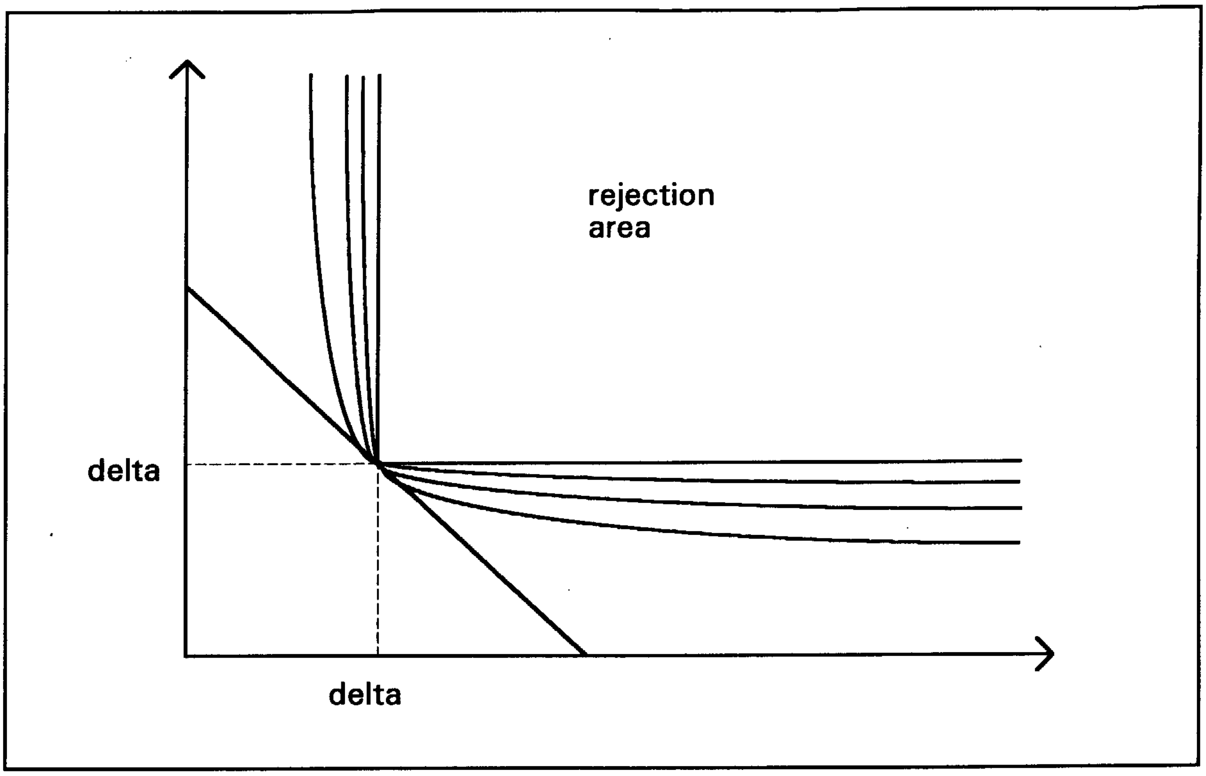

Figure 4.3.1 Symmetrical convex test-curves

All of the tests pass through the point (δ,δ). They are all symmetrical across the line $x=y$, and they are all convex.

The convex rejection areas are bounded by a cone with vertex in (δ,δ). It is geometrically evident that this cone cannot be more acute as this would imply that there would exist a dip A , not subject to rejecting, and a dip B whose both leaps were less than the ones of dip A and would be rejected.

As one can see geometrically from figure 4.3.1, the region consisting of all convex curves, symmetrical across the line x=y and intersecting the point (δ,δ), is bounded by the tests

$$\left|\frac{x_i - x_{i-1}}{t_i - t_{i-1}}\right| + \left|\frac{x_{i+1} - x_i}{t_{i+1} - t_i}\right| > 2\delta \qquad (4.3.1)$$

and

$$\min\left(\left|\frac{x_i - x_{i-1}}{t_i - t_{i-1}}\right|, \left|\frac{x_{i+1} - x_i}{t_{i+1} - t_i}\right|\right) > \delta \qquad (4.3.2)$$

As a special case one sees that the original dip-test (2.1.1) is included among these tests as a combination of (4.3.1) and (4.3.2).

Hence it is possible to find a less and a more dip-critical test than (2.1.1), namely (4.3.1) and (4.3.2). Proposition 4.3.1 summarizes the results of this section.

Proposition 4.3.1

The dip-test given in definition 2.1.1 is a member of the family of symmetrical convex dip-tests bounded by (4.3.1) and (4.3.2) in respect of geometrical rejection area.

The sharpest dip-test is given by the formula (4.3.2), and the bluntest is given by (4.3.1).

◇

When applying (4.3.1) and (4.3.2) in quality control, a pre-test $(x_i - x_{i-1})(x_i - x_{i+1}) > 0$ should be used in order to ensure that the leaps are in opposite directions.

It should also be emphasised that formula (4.3.1) is an improvement of (2.1.1) in the sense of the word that it will recognize dips in both situations described in section two, but without the drawback mentioned below these.

5. EMPIRICAL EXPERIMENTS

In this section a sample of eleven years of hourly observations from weather station 8711 Andøya, as described in [1] is used for empirical experiments. The results will be illustrated by the following three parameters:

TT	Temperature	(Degrees Celsius)
FF	Wind velocity	(knots[1])
UU	Relative humidity	(%)

In implementing the dip-tests for this exposition, formulas (3.2.1), (4.3.1) and (4.3.2) where used.

5.1 Comparing different dip-tests

In the figures 5.1.1 to 5.1.3 the relative frequency of dips recorded is shown graphically, as the delta criterion varies.

In all figures the following is apparent:

- All the curves have asymptotes for $y=0$ and $x=0$.
- The acceleration of the curves is decreasing with delta
- In the delta-area selected, the pointwise difference between the original dip-test and the "min-dip-test" (4.3.2) is greater than the difference between the original and the "sum-dip-test" (4.3.1) for small values of delta and less for greater values.

In figure 5.1.1 the curves associated with temperature (TT) is showed. In the data set there was a total of 95956 registrations, which implies that the number of dips cannot exceed 95955. As a fact, 35653 dips were observed in the data set, which is 37% of the attainable.

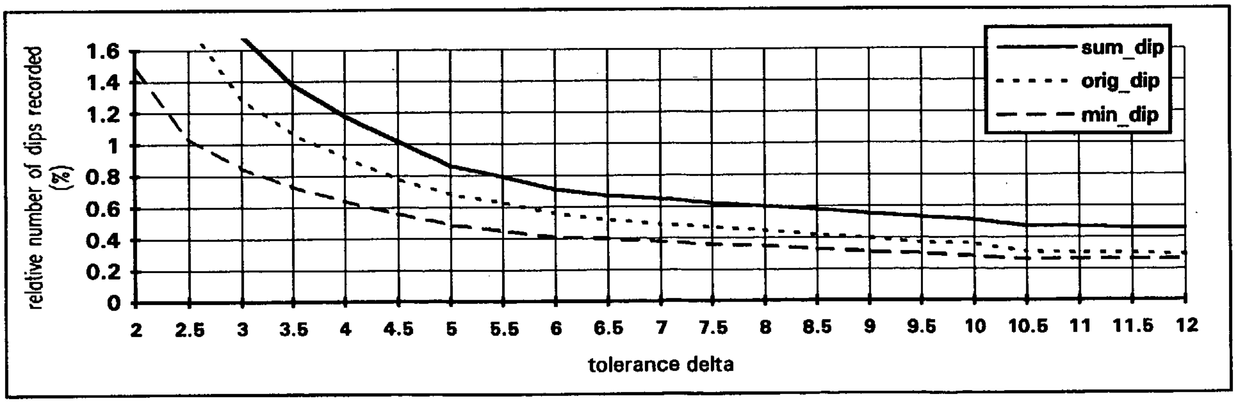

Figure 5.1.1 Relative frequency (%) of suspicious temperature values (TT) as a function of the delta-criterion.

[1]Norwegian knots, 1 knot = 0.514444 m/s, [4].

If delta is chosen equal to 5.0, as described in example 2.1.3, the number of dips observed by the sum-dip-test, the orig-dip-test and the min-dip-test is 306 (0.9%), 242 (0.7%), 171 (0.5%) correspondingly.

Figure 5.1.2 displays the curves for wind velocity. A total of 36757 dips where recorded, which amounts to 38% of the attainable. By converting the delta value specifically given for wind velocity in example 2.1.3, the corresponding delta value in knots is 14.5. The number of dips recorded with this value is 165 (0.5%), 118 (0.3%) and 77 (0.2%).

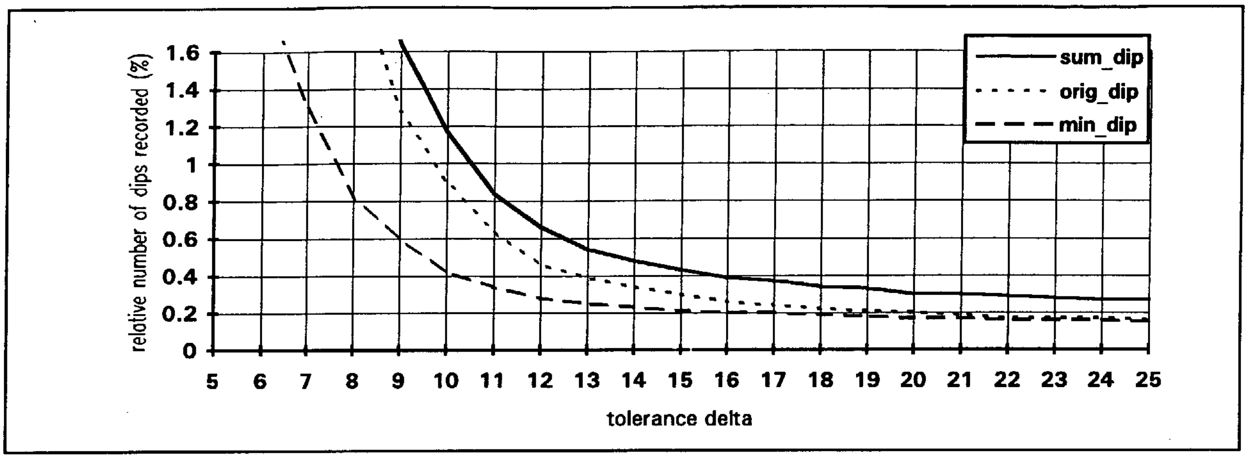

Figure 5.1.2 Relative frequency (%) of suspicious wind velocity values (FF) as a function of the delta-criterion.

Figure 5.1.3 displays the curves for relative humidity. A total of 27448 dips among the smaller population of 81939 attainable were recorded, (33%). The number of dips recorded with a delta value equal to 30.0, as proclaimed by example 2.2, is 280 (1.0%), 205 (0.8%) and 145 (0.5%).

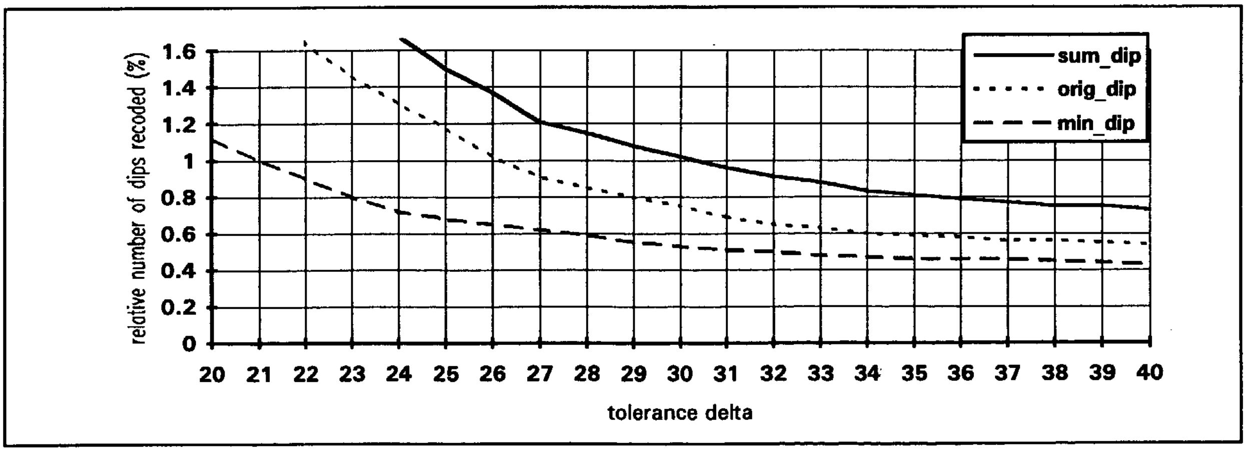

Figure 5.1.3 Relative frequency (%) of suspicious relative humidity values (UU) as a function of the delta-criterion.

5.2 Deciding good rejection criteria

To pick rejection criteria solely from statistical judgement could result in wrongly rejecting correct observations corresponding to extreme and infrequent geophysical situations. Some suggestions should however be made.

5.2.1 The original dip-test

All the values used in example 2.1.3 proved reasonable according to the graphs. As one recognises visually from figures 5.1.1 to 5.1.3, these delta values lies close to the point where the number of dips recorded by the original dip-test is the exact mean between the numbers recorded by the sum-dip-test and the min-dip-test

The number on dips recorded by the original dip-test is exactly the mean between the numbers recorded by the sum-dip-test and the min-dip-test in the following situations:

TT: δ=6.0 °C
FF: δ=13 kn.
UU: δ=27 %

5.2.2 The improved dip-test

As a way of selecting suspicious and interesting delta values for the sharper min-dip-test, the following scheme proved good in the quality control [1].

1. Make notice of the difference between the number of recorded dips for two consecutive values of delta.

2. Select the smallest value of delta from where the differences are becoming reasonably small.

As shown graphically in section 5.1, when the curves have a tendency of becoming flat, the number of dips recorded is relatively small.

An informal selection of values according to the procedure above could result in values

TT: δ=10.5 °C
FF: δ=15 kn.
UU: δ=30 %.

As these values are larger than the ones for the original dip-test and the min-dip-test is sharper than the original, a smaller number values will be recorded as suspicious. The min-dip-test focuses on dips, so values marked suspicious but not erroneous should not appear as frequently as described in example 2.2.2.

It is also possible to use delta values specified for the original dip-test (section 5.2.1) in the improved test. In this case more dips will be recorded than by using the procedure above, but still less than recorded by the original dip-test, and among the dips recorded values marked suspicious but not erroneous should not be as frequent as in example 2.2.2.

REFERENCES

[1] Øgland, P. & Førland, E. J. [1993]
DNMI-rapport nr 05/93 KLIMA "Klimastatistikk for værstasjon 8711 Andøya 1981-1991"

[2] Kjensli, P. O., Øgland, P. & Lystad, S. L. [1993]
DNMI-rapport 01/93 KLIBAS "Kontroll av data. Spesifikasjonsrapport. Klimadatabaseprosjektet. Delprosjekt 7.2"

[3] Moroney, M. J. [1951]
Facts from Figures. Penguin Books.

[4] Øgrim, O. [1975]
Størrelser og enheter i fysikken. Universitetsforlaget.