Advances in Radio Science (2004) 2: 259–261

Advances in Radio Science

Robust statistics for ionosphere data: from Rawer's nombre de corrélation (1951) to maximum-depth regression technique (1998)*

J. Taubenheim

Falkenbrunn Str. 1, D-12524 Berlin, Germany

*Dedicated to Prof. Dr. K. Rawer on the occasion of his 90th birthday.

Abstract. Contrast between "normal" and "disturbed" states of the ionosphere early induced the suggestion to present ionospheric data with the aid of a "robust" (i.e. outlier-resistant) statistic, namely, the median instead of the conventional arithmetic mean. K. Rawer, in 1951, defined on this concept a *nombre de corrélation*, which proved to be well resistant to misleading effects of outlier data when analyzing correlations between ionospheric and related data. Various attempts over many years, however, to extend this idea towards a robust fitting of regression models to outlier-contaminated data, remained unsatisfactory. Only a few years ago, a mathematically correct and unambiguous technique for robust regression has been reached by Belgian authors (Rousseeuw and Hubert, 1998), based on the new paradigm of "maximum data depth", which is exemplified here for ionospheric data presentation.

The conventional statistical techniques for presentation and analysis of data, as they are usually taught at universities, are based on the assumption that the "statistical scatter" of data follows a well-balanced, single-humped error distribution function. Its prototype is the well-known Normal Distribution established by C. F. Gauss one and a half century ago. In Geophysics, however, we have to deal with various natural systems which do not behave "normal" in this sense, but are subject to disturbances by temporary geophysical or solar-terrestrial events, producing so-called outlier data, i.e. data which do not fit into the conventional random-error distribution model. Application of conventional statistical procedures to data which are "contaminated" by outliers can lead to erroneous interpretations and wrong conclusions. Just in the ionosphere, we typically have always to live with contaminated data of this kind, forming indeed a minority but nevertheless a physically relevant part of our empirical material. Consequently, the fathers of world-wide ionospheric research were wise enough to recommend the use of statistics which are sufficiently resistant to outlier data, e.g. medians instead of arithmetic means, and quartile ranges instead of r.m.s. deviations. In statistical theory, characteristics which in this way are insensitive to deviations from a basic distribution model are called robust.

Correspondence to: J. Taubenheim
(jens.taubenheim@t-online.de)

Heavy and unwanted misleading effects can happen when outlier data are involved in a study of correlations of various ionospheric data between each other or with other atmospheric or solar predictors. To exemplify this, Fig. 1 (see next page) shows a plot of daily data of night-time F-region electron density in dependence on a solar activity index. (The sample is conveniently selected for its tutorial purpose, but consists of real data obtained at our ionosonde station Juliusruh/Rügen). Its visual impression suggests an obvious tendency of increasing ionisation with increasing solar activity, just as we should expect it. With high solar activity, however, there happen ionospheric storm events which drastically reduce the electron density. In effect, when we apply the classical regression procedure of least-square fitting we do not obtain a significant positive correlation between ionosphere and solar activity, but seemingly even an indication of slight negative correlation! Clearly, the shape of the "scatter cloud" of observed data points is not adequately represented by the conventionally calculated regression line. This is caused by the fact that the classical algorithm of minimizing the sum of squared residuals damages itself since it puts a heavy overweight just on the outliers with their large residual values!

It was an idea of Rawer (1951), already half a century ago, to take advantage of the robust properties of the median by designing an alternative technique for the test of correlations, in order to avoid the misleading effects of outlier data. He published it as an annex to a paper in which he compared observations made at two stations: The scatterplot of the two variables is divided by their medians into four quadrants, A, B, C and D (see Fig. 2). If the variables were really uncorrelated, we evidently have to expect, by virtue of the definition of medians, that an equal share of the total number of observations should fall into each of the four quadrants. On the other hand, if there exists a (positive or negative)

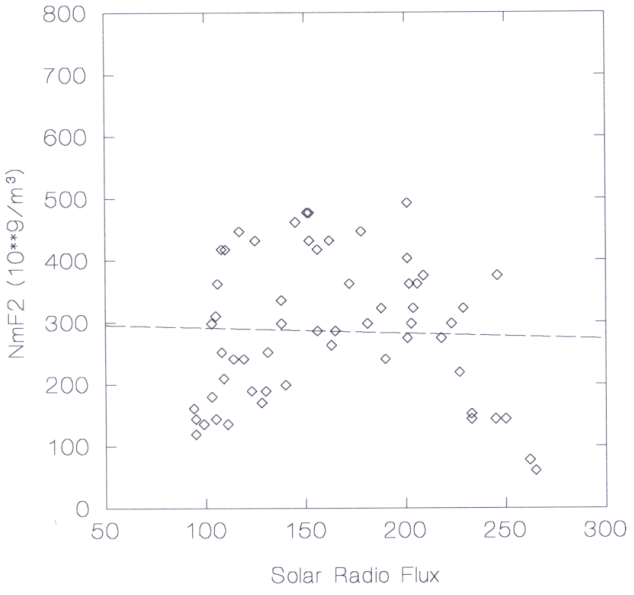

Fig. 1. Pre-dawn minimum of F2-layer peak electron density (NmF2), versus solar 10.7 cm radio emission. Daily values in June–July, 1982, at Juliusruh/Rügen ionosonde station. Straight line: Linear fit according to conventional least-square regression.The Pearson correlation coefficient is virtually zero.

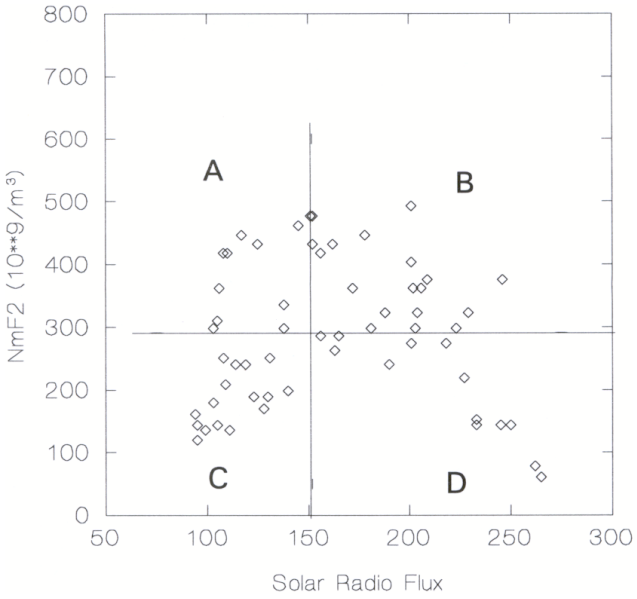

Fig. 2. Same data as in Fig. 1, but divided into quadrants, according to Rawer (1951).The number of points in quadrants B + C is greater than in A + D, indicating a positive correlation.

correlation, this would express itself by the number of points in $B$and C being significantly greater (or smaller, respectively) than those in $A$and D. Accordingly, Rawer defined a statistical measure, called by him *nombre de corrélation*, to be calculated from the crosswise ratio of the "point mass" in $B + C$to that in $A + D$.In our example (Fig. 2) this ratio is $32/26 > 1$, clearly indicating a positive correlation.

A few years later, I proposed a slight amendment of this definition, facilitating its smooth transformation into the conventional Pearson correlation coefficient when the observed data approach the "classical" two-dimensional normal distribution (Taubenheim, 1958).

It can be easily understood, that this "quadrant technique" proposed by Rawer provides a measure of correlation which is well outlier-resistant, i.e. "robust", for the analysis of disturbance-contaminated ionospheric data. Moreover, it is quick and easy to perform, and permits a Chi-square-test of significance according to the usual standards of statistical analysis.

When we analyze correlations, however, our final aim is not only to prove the existence of a correlation between observed quantities, but to incorporate it in predictive models. That means we have to strive for a regression algorithm which is robust as well, in order to be safe of unwanted effects like that found in our example. Curiously enough, this task turned out to be more difficult than expected. Various attempts to it were published and discussed in the course of the past decades, but in general they proved not fully satisfactory under the demands of mathematical strictness and numerical practicability. Only very recently, a few years ago, a solution of this important problem has come up, which appears both mathematically correct and well practicable. Here I cannot go into the details of the theory behind it, but I want to draw in brief your attention to it.

The starting-point is a new statistical paradigm of data depth, which has been introduced in the late 1990s by American statisticians (see, e.g. Liu et al., 1999). In brief, the notion of "depth" means how "deeply embedded" is a point in the scatter cloud of a given n-dimensional data sample (in other words, how comprehensively this point is "surrounded" by the given data cloud). After this definition, it is immediately plausible that the point located at maximum depth in a 1-dimensional sample is just our well-known median. Similarly, there exists a point of deepest location in any 2-, 3-, or higher-dimensional sample as well. Data depth is a robust statistic, i.e. it is resistant to outliers and insensitive to assumptions about the shape of the distribution functions of the variables.

For the problem of robust correlation and regression, a breakthrough came when Rousseeuw and Hubert (1999), working at the University of Antwerp, extended the paradigm of data depth by defining a robust "regression depth", which in a similar sense characterizes how deep a regression line is "embedded" (or "nested") in the scatter cloud of a data sample. Consequently, the best regression fit is then ascribed to a line which is located at maximum regression depth in the scatter cloud. This is the line which takes, so to say, the "most balanced" position within the surrounding data. The authors proved that the paradigm of regression depth can not only be applied to 2-dimensional $y(x)$ regression, but to higher dimensional regression as well.

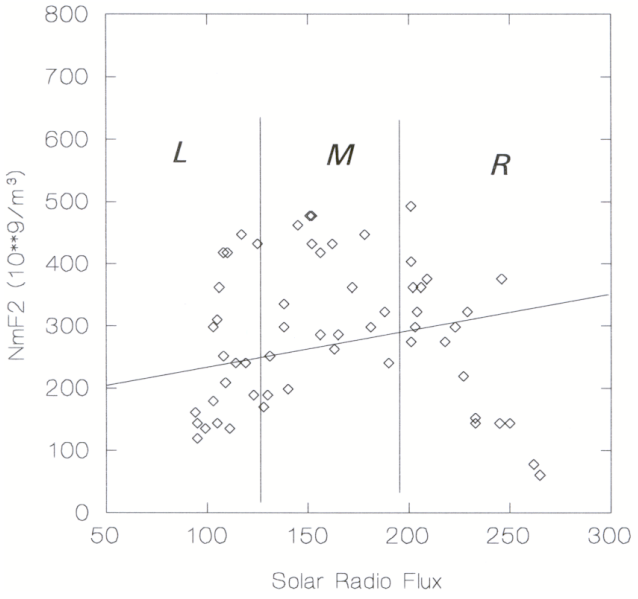

Fig. 3. Same data as in Fig. 1, but divided into thirds L, M, R, according to Hubert and Rousseeuw (1998). The straight CAT-line which "cuts all 3 thirds" represents maximum "regression depth".

For the construction of a regression line which well approximates maximum depth in a 2-dimensional scatter plot, Hubert and Rousseeuw (1998) have given a rather efficient algorithm, which can be outlined here in short as follows (see Fig. 3): Divide the N data of the sample along the abscissa into three equal parts, of $N/3$ data points each, labelled "Left" (L), "Middle" (M), and "Right" (R), and draw a first tentative (zero-order) straight regression line which cuts the scatterplot into 3×2 fields, three above (+) and three below (-) this line. Then improve the position of the regression line by an iterative procedure, until finally in each of the combined fields $L^+ \cup M^+$, $M^+ \cup R^+$, $L^- \cup M^-$,and $M^- \cup R^-$, an equal number of $N/3$ data points is found. The iterative procedure in general converges rather rapidly. The resulting regression line is called the "CAT-Line", because it cuts all 3 thirds of the sample. It can be proven that the CAT-line takes the "deepest possible" position of a straight line in the sample. Obviously, it represents a (robust) 2-dimensional generalization of the (1-dimensional) median.

We see that the CAT-line in our example (Fig. 3) characterizes the correlation between the ionosphere and solar activity in a much more plausible way than with the classical least-square regression displayed in Fig. 1. As we have already recognized that robust statistics are the appropriate tool for the analysis and empirical modeling of disturbed ionospheric data, we can now state that with this robust regression a long-wanted progress has been achieved on a way which was visualized by K. Rawer several decades ago.

References

Hubert, M. and Rousseeuw P. J.: The Catline for deep regression, J. Multivariate Anal., 66, 270, 1998.

Liu, R. Y., Parelius, J. M., and Singh, K.: Multivariate analysis by data depth, Annals of Statist., 27, 783, 1999.

Rawer K.: Comparaison des résultats de mesures ..., J. Atmos. Terrest. Phys., 2, 38, 1951.

Rousseeuw, P. J. and Hubert, M.: Regression depth, J. Amer. Statist. Assoc., 94, 388, 1999.

Taubenheim, J.: Ein einfaches Korrelationsmaß, Naturwissenschaften, 45, 413, 1958.