SCIENTIFIC PUBLICATIONS OF THE STATE UNIVERSITY OF NOVI PAZAR
SER. A: APPL. MATH. INFORM. AND MECH. vol. 14 (2) 2022, 73-79

The Statistical Behavior of Dust-related Radio Waves

K. Racković Babić

Abstract: Radio and plasma wave instruments in space can detect cosmic dust over a wide range of sizes through impact ionization. Our understanding of dust particle properties can be gained from analyzing such electric pulses. In order to explain how dust particles produce electrical signals, several physical mechanisms have been proposed. Recently, Rackovic Babic et al. (2022) developed a model which takes into account all the effects of charge collection by the spacecraft and electrostatic influence from charges in its vicinity. The authors tested the model's accuracy using the database specially created for a given purpose. This paper aims to examine the mathematical significance of the obtained results from the fitting method.

Keywords: data analysis, modeling, statistics

1 Introduction

The dust impacts on spacecraft produce measurable electrical signals. Such transient voltage signals are generated by the expanding plasma cloud after impact ionization. The antenna instruments can measure these voltage signals that provide information about dust particles. Hence, developing models of how signals are generated is important in order to be able to link observed electric signals to the physical properties of the impacting dust. The obtained parameters from fitting the model to measured waveforms can provide information on dust particles and characteristics of impact-generated plasma cloud, as well as characteristics of ambient plasma environment. Several models have attempted to describe the physical mechanisms leading to the generation of voltage signals measured by antennas (e.g. [7],[3]). In this paper, we will concentrate on the model proposed by [5], which takes impact-ionization-charge collection and electrostatic-influence effects into account.
The accuracy of the model was tested using dust-related data from the STEREO satellite. The results obtained provide insight into interesting physical phenomena related to dust impact. We tend to evaluate the proposed model's accuracy through a mathematical approach in the present work.

Manuscript received 23 August, 2022. ; accepted October 31, 2022.
K. Racković Babić is with the Faculty of Mathematics, Univeristy of Belgrade, Serbia

2 Fitting model

The model will be briefly presented below. Physic backgrounds are out of the scope of current work, so we won't discuss it here. A detailed description of the model can be found in the paper [5]. Function

$$\phi(t) = A(1/(1-T_1/T_2))(e^{(-t/T_2)} - e^{(t/T_1)}) - Be^{(-t/T_2)}, \tag{1}$$

with four free parameters (A, B, $T1$ and $T2$) is the core of our interest. The function was used to fit the signal recorded by the STEREO/TDS instrument [1], and a Levenberg-Marquardt least-squares minimization method was used to accomplish this. Such a procedure produced very interesting empirical results. We will now use the same function, and the same data set in order to test the hypothesis statistically. By determining the odds that obtained results occurred by chance, we are determining whether obtained results are valid.

3 Analysis

The previous research experience offered the above relation (Eq.(1) for describing our events. This means that we know the functional relationship between the measurement results and only need to determine the coefficients. When the statistical conditions are met, the least-squares method (LSM) can be used to determine those values. For LSM to be entirely reliable, it should be possible to subsume our relation into one of the acceptable models, which will not be possible with Eq.(1). Therefore, the statistical modeling was performed using the Levenberg-Marquardt method (LM). Our goal here is to evaluate the quality of the performed modeling.

The following quantities can describe a real object (system, process, phenomenon, etc.): input variables (X) - describe the operating (in the mathematical models known as independent variables); output variables (Y)- provide behavior or the result (known as dependent variables); hidden variables (random residuals) (E) - cannot be directly measured, but show influence on the dependent variable of factors that are not to be taken into account at the "entrance" (known as residues).

Particularly, for the results l obtained when measuring an object $\mathscr{O}$

$$\mathscr{O} \longleftrightarrow \{(x_k^{(1)}, x_k^{(2)}, ..., x_k^{(n)}, y_k^{(1)}, y_k^{(2)}, ..., y_k^{(m)})\}, \quad (k = \overline{1,l}) \tag{2}$$

using Eq.(1), we created the transformation

$$\vec{f}\{\mathscr{O}\} = \vec{f}(x^{(1)}, x^{(2)}, ..., x^{(n)}) = \left\{ \begin{array}{c} f^{(1)}(x^{(1)}, x^{(2)}, ..., x^{(n)}) \\ f^{(2)}(x^{(1)}, x^{(2)}, ..., x^{(n)}) \\ \vdots \\ f^{(m)}(x^{(1)}, x^{(2)}, ..., x^{(n)}) \end{array} \right\} \tag{3}$$

which should enable the best calculation of the dependent variables Y from the set values of the independent variables X. There is a difference in meaning and designations between the sample mean and conditionally theoretical mean, $\bar{y}(x)$ i $y_m(x)$, and since we do

not know the absolute measurement error of a quantity, in practice we also do not know the theoretical mean, but we rely on the fact that, according to the law of large numbers $\bar{y}(x) \xrightarrow[\text{by probability}]{} y_m(x)$, when the number of observations, from which $\bar{y}(x)$ is calculated, tends to infinity.

In the examples, we will denote theoretical averaging with **E** and dispersion of the random variables with **D**. The best-based and most accurate solution relies on probabilistic knowledge - ie. the form of the distributive law is likely to be the residual ε in the selected model. Assuming the normal law gives the probability distribution of the residual ε, with parameters $(0,\sigma^2)$ for an arbitrary realization x, where σ is a constant, and $[\varepsilon(x_i),\ i=1,n]$ are mutually statistically independent, the lowest error for $y_m(x)$ according to model $f(x) \subset \vec{f}(x)$ can be obtained by the LSM as $\Delta_n(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 \to \min_{f\in\vec{f}}$.

Different residue types may permit different methods of selecting the quality of approximation Δ_n. Anyway, we analyze the relation between y and $(x^{(1)},x^{(2)},...,x^{(n)})$, in the form $y = f(x^{(1)},x^{(2)},...,x^{(n)};\Theta) + \varepsilon$, where ε is residue and $f(X;\Theta)$ is the function of some known parametric family $\vec{F} = \{f(X;\Theta)\}$, $\Theta \subset A$ into which an unknown numerical value of the parameter Θ is inserted. We do not consider the analytical form of $Y_m(X) = f(X;\Theta)$ here, but rather the nature of the variables analyzed (X,y), as well as their interpretation of the function $f(X;\Theta)$. Thus, if there is a dependence of the value of the resulting characteristic η not only on the value of X, but also on uncontrolled factors such that for each fixed value of (X^*), the corresponding values of the random characteristic are $\eta(X^*) = (\eta|X = X^*)$ exposed to some random scattering, we can evaluate the quality of our approximation from that feature. Here, the predictor variables X are in the role of a non-random parameter on which the probability distribution law - especially the mean value and dispersion - of the investigated resulting criteria η depends. The following mathematical model is suitable for dependencies of this type

$$\eta(X) = f(X) + \varepsilon(X), \tag{4}$$

where $f(X)$ describes the behavior of the conditional mean $y_m(X) = \mathbf{E}\eta(X) = f(X)$ based on X, and the remaining component $\varepsilon(X)$ is a reflection of the nature of $\eta(X)$ [2]. Note here, $\mathbf{E}\varepsilon(X) \equiv 0$ based on the assumption that for all X there is a finite dispersion for $\varepsilon(X)$ ($\mathbf{D}\varepsilon(X) < \infty$), while the dispersion may depend on X ($\mathbf{D}\varepsilon(X) = \sigma^2(X)$). Note that in the model, neither the random component's nature nor its probability distribution are related to the structure of $f(X)$ and, in particular, are not dependent on parameter Θ in the parametric model form - when instead of all functions $f(X)$ considers any $f(X;\Theta)$. In this way, the permissible nature compositional parts of the model (3) are concretized: $y_m(x) = \mathbf{E}\eta(x) = f(x) = \theta_0 + \theta_1 x$ and $\sigma^2(x) = \mathbf{D}\varepsilon(x) = \sigma_0^2(y_m(x))^2$, where σ_0 is constant. Typically, both vectors, η i ξ, are influenced by numerous uncontrolled and random factors. Thus, it is useful to split the resulting feature η into two random components

$$\eta = \mathbf{f}(\xi) + \varepsilon, \tag{5}$$

where the left-hand side part is determined by vectorized function **f** of the variables **f**, and ε is the remaining component, as long as the components of the vector $\mathbf{f}(\xi)$ and ε meet

the conditions: $\mathbf{E}\varepsilon^{(k)} = 0, \quad \mathbf{D}\varepsilon^{(k)} = \sigma_k^2 < \infty, \quad cov(f^{(k)}(\xi), \varepsilon^{(k)}) = \mathbf{E}|(f^{(k)}(\xi)\varepsilon^{(k)})| - \mathbf{E}f^{(k)}(\xi)\mathbf{E}\varepsilon^{(k)} = 0$. For a unique resulting feature ($m = 1$) and a linear expansion of $f(\xi)$, we have:

$$\eta = \theta_0 + \sum_{k=1}^{p} \theta_k x^{(k)} + \varepsilon. \tag{6}$$

Assuming $y_m(X) = \mathbf{E}(\eta|\xi = X)$, Eq.(6) turns into a linear regression equation $y_m(X) = \theta_0 + \sum_{k=1}^{p} \theta_k x^{(k)}$. It is possible that ε in the Eq.(4) is most likely (with a probability of 1) equal to zero. Then η and ξ are only connected by $\eta = f(\xi)$ (this should not be confused with the functional connection of non-random variables).

A class of admissible solutions F should be selected based on the analysis of the character and degree of statistical relationships between the examined variables, i.e., the parametric family of functions $F(X)$, from which the best approximation $< \mathbf{f}(X) >$ of the required dependence is selected. This aims to determine which approximations represent the best solution to the shape optimization

$$< \mathbf{f}(X) > = \arg \underset{\mathbf{f} \in \mathbf{F}}{\operatorname{extr}} \Delta_n(\mathbf{f}) \ , \tag{7}$$

Where the functional $\Delta_n(\mathbf{f})$ determines the quality criterion of the resulting approximation, which will be denoted by η (or Y) by the function $f(X)$ from class F. The choice of the specific form of that functional relies on the knowledge of the probabilistic nature of the residual ε in the selected models (Eq.(4), Eq.(5) and Eq.(7)). The LSM functional is the most widely used. If parametric families of functions $\mathbf{f}(X;\Theta)$ are given in the property of class $\mathbf{f}$, the analysis becomes statistical, parameter values $< \Theta >$ for which the maximum is reached by Θ of the functional $\Delta_n(\mathbf{f}(X;\Theta))$ and corresponding models are called parametric.

Last but not least, the phase of analyzing the accuracy of the connection equations confirms that the approximation $< \mathbf{f}(X) >$ of the unknown theoretical function $\mathbf{f}_T(X)$, which was found in accordance with Eq.(1) and based on the Eq.(3), Eq.(4) and Eq.(6), represents only an approximate representation of the real dependence $\mathbf{f}(X)$. The error δ of the description of the function $\mathbf{f}_T(X)$, via $< \mathbf{f}(X) >$ in the general case has two components of: approximation errors ($\delta_{\mathbf{F}}$), and errors due to sampling ($\delta(n)$). The value of $\delta_{\mathbf{F}}$ is correlating with the choice of class $\mathbf{F}$. When $\mathbf{f}_T(X) \in \mathbf{F}$ (special cases), $\delta_{\mathbf{F}} = 0$. Additionally, $\delta(n)$ remains due to the sample's own limitations. This error can be reduced by increasing the sample size (n). Thus, for $\delta_{\mathbf{F}} = 0$ and with correctly chosen methods of statistical evaluation sampling, error $\delta(n) \to 0$ when $n \to \infty$ by probability.

This phase is illustrated graphically in Figure 1. The figure illustrates well the need to apply the above theoretical operations to test the possibility of eventual improvement of the initial function (Eq. (1)). The total number of dust-related pulses in our dataset is 349632. This means we are referring to the so-called a posteriori correction of theoretically, experientially, and intuitively obtained relations, while this correction must be performed statistically. Due to the sample size, we leave this task to the next phase of our research and explain the necessity of such steps. The theoretical basis for this correction is presented

below.

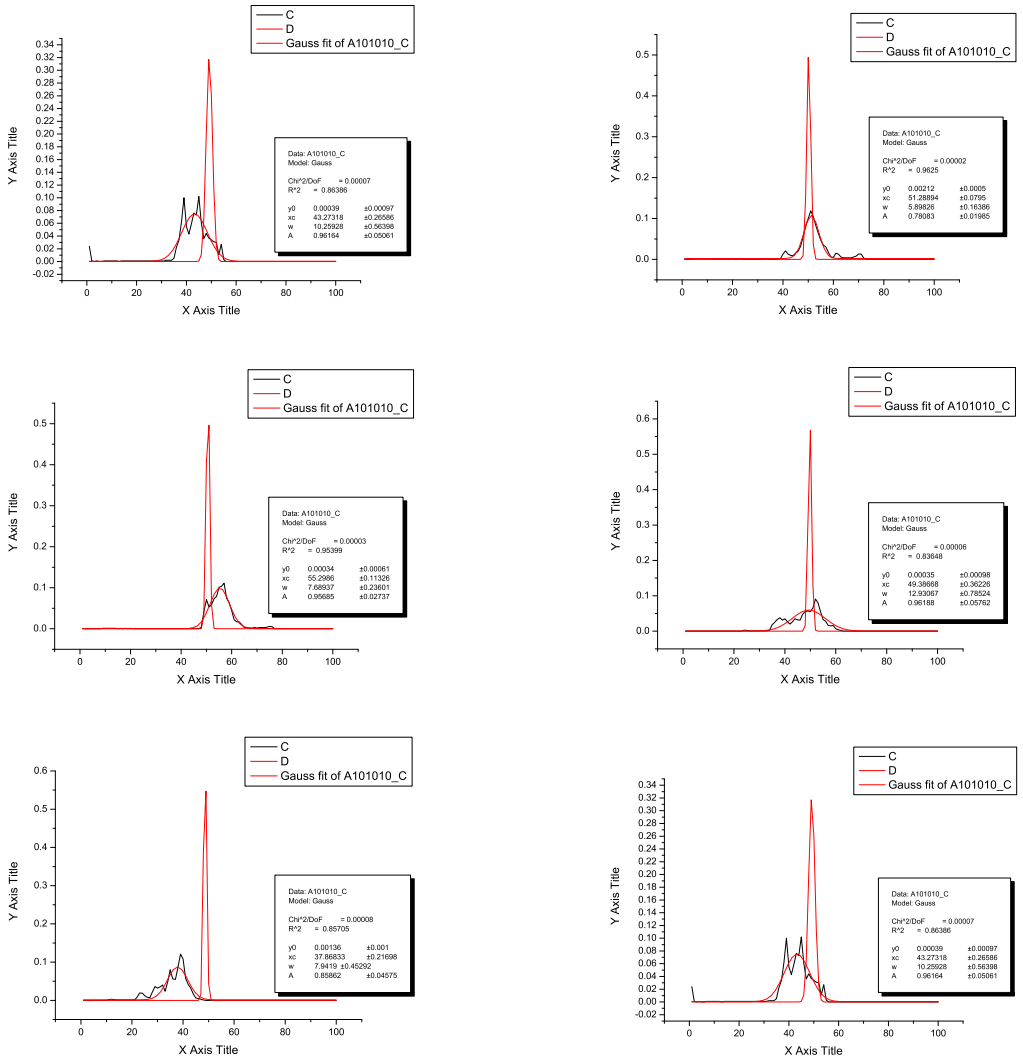

Fig. 1. Residue behavior in the X, Y, and Z antenna system, with two events given for each monopole. It is evident that each monopole behaves as a part of the system, which in its entirety provides a normal distribution for the residuals, but with visible excesses and moments. In each individual image, the frequency distribution of the residuals in the broader environment of the mean (it should be 0) is indicated in black color, as well as the corresponding approximation with a Gaussian curve. The curve shown in red with a pronounced peak is a theoretical Gaussian curve with the parameters of the mean and the dispersion of the residual. It is usually centered and zero outside the $\pm 3\sigma$ zone.

A qualitative criterion must be specified based on which we can select the best predictive model $y_m(x)$. The best solution is determined by knowing the probabilistic nature - and thus

the law of probability distribution - of the residual ε in the selected model. This criterion is provided by the LSM under certain conditions.
It is explicitly incorrect in our example that the residual dispersion is constant, which means that the conditional dispersion $\mathbf{D}(\varepsilon|\xi = x) = \mathbf{D}(\eta - \theta_0 - \theta_1 \cdot \xi|\xi = x) = \sigma^2(x)$ actually depends on x.
Assume we divide all dependent variables by dispersion $s(x)$. Now we have some sort of normalized residual $\tilde{\varepsilon}(x) = \varepsilon(x)/s(x)$. At a later stage, we will demonstrate that in such a situation, the assumption that the distribution function is likely to be $F(\tilde{\varepsilon}) = \mathcal{N}(0, \sigma^2)$, $\sigma \neq \sigma(x)$x) is valid. The previous minimization leads to the task of determining the extremum $\Delta_n(f) = \Delta_n(\theta_0, \theta_1) = \sum_{i=1}^{n} \left(\frac{y_i - \theta_0 - \theta_1 x_i}{s(x_i)}\right)^2 \to \min\limits_{\theta_0, \theta_1}$, specifically, to the system:

$$\frac{\partial \Delta_n(\theta_0, \theta_1)}{\partial \theta_0} = -2 \sum_{i=1}^{n} s^{-2}(x_i) \cdot (y_i - \theta_0 - \theta_1 x_i) = 0$$

$$\frac{\partial \Delta_n(\theta_0, \theta_1)}{\partial \theta_1} = -2 \sum_{i=1}^{n} s^{-2}(x_i) \cdot x_i \cdot (y_i - \theta_0 - \theta_1 x_i) = 0.$$

As a result of solving this system, we get the estimations $\widehat{\theta}_0$ and $\widehat{\theta}_1$ of the unknown parameters θ_0, θ_1 via the system

$$\widehat{\theta}_0 = \frac{\left[\sum\limits_{i=1}^{n} s^{-2}(x_i)\right]\left[\sum\limits_{i=1}^{n} s^{-2}(x_i) \cdot x_i y_i\right] - \left[\sum\limits_{i=1}^{n} s^{-2}(x_i) \cdot x_i\right]\left[\sum\limits_{i=1}^{n} s^{-2}(x_i) \cdot y_i\right]}{\left[\sum\limits_{i=1}^{n} s^{-2}(x_i) \cdot x_i\right]^2 - \left[\sum\limits_{i=1}^{n} s^{-2}(x_i)\right]\left[\sum\limits_{i=1}^{n} s^{-2}(x_i) \cdot x_i^2\right]}$$

$$\widehat{\theta}_1 = \frac{\left[\sum\limits_{i=1}^{n} s^{-2}(x_i)\right]\left[\sum\limits_{i=1}^{n} s^{-2}(x_i) \cdot x_i y_i\right] - \left[\sum\limits_{i=1}^{n} s^{-2}(x_i) \cdot x_i\right]\left[\sum\limits_{i=1}^{n} s^{-2} \cdot x_i y_i\right]}{\left[\sum\limits_{i=1}^{n} s^{-2}(x_i)\right]\left[\sum\limits_{i=1}^{n} s^{-2}(x_i) \cdot x_i^2\right] - \left[\sum\limits_{i=1}^{n} s^{-2}(x_i) \cdot x_i\right]^2}$$

4 Discussion and Conclusion

In parametric analysis, even when experience indicates the model type, the parameters of that model can vary significantly in accuracy, while transformations of translation and rotation of the resulting regression curves represent the most straightforward cases that can already be identified a priory at the level of experience. Suppose the considered parameter within the general population has a normal distribution. In that case, the mathematical expectation and dispersion (mean-square deviation) are sufficient to define that distribution, i.e., evaluates those parameters because they entirely determine the normal distribution. Therefore, the statistical evaluation of the unknown parameter of the theoretical distribution is a function of the experimental (sampled) random variables. Being aware that the use of a sample to estimate parameters could lead to incorrect estimation of those parameters,

here we list the basic requirements that the estimation of such parameters must meet to arrive at a reasonable value ultimately.
Condition $E(\widehat{\Theta}) = \Theta$ will not remove errors in the values of individual assessments, but it will eliminate systematic errors in assessments. Known as the centrality of the assessment, this condition does not depend on sample size. Otherwise, the assessment is not centered. Note the centrality of the rating does not guarantee the right to accept every centered rating as reliable. Hence, another characteristic of statistical assessment must be introduced - the effectiveness. A statistical assessment is effective for a given sample if it has the smallest possible dispersion $\lim_{m \ll m_{max}} D(\widehat{\Theta}_i^{(m)}) \longrightarrow min$. However, suppose an evaluation of a parameter is "sensitive" to a change in the sample volume. In that case, the question of its stability is also raised, so the third requirement for statistical evaluations with large sample volumes is stability. An estimate $\widehat{\Theta}$ of the parameter $\widehat{\Theta}$ is considered stable in probability if $\lim_{m \to \infty} \widehat{\Theta}_i^{(m)} \xrightarrow{v} \Theta, \;\; i = 1, n$.
Based on the preliminary analysis of our data, it seems necessary to carry out a comprehensive analysis of experimental data in terms of modeling all residuals, as a whole and by the event.

References

[1] J. L. BOUGERET, K. GOETZ,M. L. KAISER,. ET AL., *S/WAVES: The Radio and Plasma Wave Investigation on the STEREO Mission* Space Science Reviews 136 (2008) 487–528.

[2] D. DJUROVIC, *Matematicka obrada astronomskih posmatranja*, Univerzitet u Beogradu, Beograd, 1979.

[3] N. MEYER-VERNET, M. MAKSIMOVIC, A. CZECHOWSKI, ET AL., *Dust Detection by the Wave Instrument on STEREO: Nanoparticles Picked up by the Solar Wind?*, Solar Physics 256 (2009) 463–474.

[4] N. MEYER-VERNET, M. MONCUQUET, K. ISSAUTIER, P. SCHIPPERS, *Frequency range of dust detection in space with radio and plasma wave receivers: Theory and application to interplanetary nanodust impacts on Cassini*, Journal of Geophysical Research (Space Physics) 122 (2017) 8–22.

[5] K. RACKOVIC BABIC, A. ZASLAVSKY, K. ISSAUTIER, N. MEYER-VERNET, D. ONIC, *An analytical model for dust impact voltage signals and its application to STEREO/WAVES data*, Astronomy and Astrophysics, A&A 659, A15 (2022)

[6] A. ZASLAVSKY, ET.AL, *Interplanetary dust detection by radio antennas: Mass calibration and fluxes measured by STEREO/WAVES*, Journal of Geophysical Research (Space Physics) 117 (2012) A05102.

[7] A. ZASLAVSKY, *Floating potential perturbations due to micrometeoroid impacts: Theory and application to S/WAVES data*, Journal of Geophysical Research (Space Physics) 120 (2015) 855–867.