Vasilios Zarikas - Theofilos Chrysikos - Konstantinos E. Anagnostou - Stavros Kotsopoulos
Panagiotis Avlakiotis - Charalambos Liolios - Theodoros Latsos - Georgios Perantzakis
Athanasios Lygdis - Dimitirios Antoniou - Asimakis Lykourgiotis *

WIRELESS TELEMETRY: CHANNEL CHARACTERIZATION AND STATISTICAL IMPUTATION OF MISSING VALUES

An implemented communication wireless telemetric system is presented. It has been built to serve as a measuring station at Thermopiles hot springs. The present work analyzes the method applying several models for adjusting and optimizing the functions of the system. The system also includes a unit that receives and stores data in an appropriate form, ready to be used for statistics. Data of critical physicochemical parameters are continuously measured, processed and transmitted over a wireless link. Path loss is investigated in order to ensure reliable signal reception. In the examined case study of low SNR region with the integrated in the open field sensors, missing data can be a common occurrence. The development of a mathematical technique, based on best prediction, for mitigating the missing values problem is of great significance in establishing a reliable system of wireless telemetry. A novel technique for handling missing values is proposed. An algorithm that evaluates missing values for imputation has been developed, based on a new general linear model analysis for prediction. The full detailed code, is explained and applied on the data that have been captured by the measuring station.

Keywords: *Telemetric communication system, path loss, measuring station, hydrogeology engineering, seismic activity, modeling/statistics.*

1. Introduction

In general, a wireless network with sensors is a system consisting of spatially distributed autonomous devices using sensors to cooperatively monitor a variety of context conditions, such as temperature, vibration/frequencies, acceleration/pressures, fields, pollutants, at different locations. Critical characteristics of a wireless network with sensors are: the size scale of sensor nodes, harsh context conditions, mobility/portability, network topology, communication failures, adjustability of dynamic operationand large scale of deployment. Such systems with sensors can be used in a variety of contexts of use; environmental monitoring, medical monitoring, acoustic/fields detection, military surveillance, or process monitoring.

In large surveys that carry on for extended time periods, it is very common to encounter the missing values problem. Measuring stations like the one under discussion, almost always comprise missing measurements. There is a variety of reasons that cause this problem, including extreme weather conditions, information system failures, power shutdowns, rare and unpredictable events, malicious acts, sensor blockages etc. Most of these events appear randomly contrary to machine failures that appear more often at the early stages of the measurements.

The proper treatment of missing values helps to address several concerns resulted by incomplete data. If there are cases with missing values which are systematically different from cases without missing values, the statistical inferences can be misleading. Moreover, missing data often decrease the precision of calculated statistical outputs since there is less information than originally designed. One more worry is that the assumptions behind many statistical procedures are based on complete cases, and missing values can complicate the theory required.

2. State of the art

In this work a measuring station is described comprising sensors for different physicochemical parameters of thermal waters (for a review of similar measuring stations see [1]). The station has been implemented for a particular hot spring located in Thermopiles, Greece. The factors that are continuously measured, using the constructed wireless network with sensors, are the concentration levels of radon in water, the water temperature to capture the flow rate and depth variations, PH to investigate the acidity variations, Redox potential to study the biologic load variations, and electrical conductivity as a measure of the salinity variations.

* [1]Vasilios Zarikas, [2]Theofilos Chrysikos, [1]Konstantinos E. Anagnostou, [2]Stavros Kotsopoulos, [1]Panagiotis Avlakiotis, [1]Charalambos Liolios
[1]Theodoros Latsos, [1]Georgios Perantzakis, [1]Athanasios Lygdis, [1]Dimitirios Antoniou, [2]Asimakis Lykourgiotis
[1]ATEI of Central Greece, Department of Electrical Engineering, Lamia, Greece
[2]University of Patras, Department of Electrical & Computer Engineering, Patras, Greece
E-mail: vzarikas@teilam.gr; kotsop@ece.upatras.gr

This study reports a) the models that have been used in order to tune and optimize the telemetric module of the measuring station, b) a novel statistical technique for handling missing values of the transmitted data, based on a new best for prediction general linear model technique and c) a detailed algorithm that realizes this mathematical method. In Appendix A the algorithm is described in pseudocode. .

The measuring station has been designed and implemented under the framework of a research project; it is a major research program that has received so far two national and EU research grants. The whole project has as a main scope to design, develop, test and optimize a novel system that integrates hardware and software modules capable to perform the required environmental measurements. It is also able to collect, maintain, transmit receive and store data. A second objective is to perform a modern statistical analysis of the data in order to understand their structure and capture the encoded information. Towards this scope we have developed, for the first time, an algorithm which is able to calculate the missing values for imputation based on a new best fitting polynomial designed to be the best as far as prediction is concerned.

A quite general architecture has been designed for the implemented platform. This platform includes also an independent source of energy via photovoltaic elements and a power management electronic module, see Fig. 1.

Fig. 1. The telemetric measuring station

In all telemetric measuring stations a crucial concern is how to minimize and consequently how to care missing values. The presented measuring station is an integrated system that was designed and adjusted after three evaluations in a way to minimize missing values for all the measuring factors from the sensors. However, it is always unavoidable such lost measurements to occur. Thus, it is important to fill in missing values. Common techniques that help impute missing values with estimated values are regression models or EM methods. The EM method assumes a distribution for the partially missing data and extract inferences on the likelihood under that distribution. Each iteration consists of an “E” step and an “M” step. The “E” step calculates the conditional expectation of the complete-data log likelihood given the observed data and the parameter estimates. In the “M” step, maximum likelihood estimates of the parameters are computed as though the missing data had been filled in. However note that the missing values are not being directly imputed. Instead, their functions are used in the log-likelihood.

The regression method instead computes multiple linear regression estimates. For every predicted case, the algorithms can even add a residual from a randomly selected complete case, a random normal deviate, or a random deviate from the t distribution. However this method and all its versions adopt the usual statistical criteria in model selection [2] which are designed to the target: find the model, wh ich “best”, under some distance criteria, fits the data. Since these criteria are, in principle, functions of the residual sum of squares [3] they can not address the need for the best prediction of the missing value. Nevertheless, these best fitting models are applied for prediction too, although these “distance” criteria were not designed for this purpose.

3. Telemetric system and Outdoor Path Loss Models

The data collection measuring station transmits the digital data through a wireless radio network (using the radio modem). The receiver main unit for data processing was placed in the campus of Technological Educational Institute (ATEI) of Central Greece, Lamia, Greece.

The wireless link between transmitter and receiver is considered to be Line of Sight (LOS). We assume that the transmitter collects all sensor data from a local wireless sensor network protocol (ZigBee) and then proceeds to transmit all data to the main receiver at the campus of ATEI in Lamia.

The status of the wireless channel characterization was studied on theoretical basis and the theoretical reliability was tested by performing a scenario of outdoor measurements using the portable broadband measuring device Narda Selective Radiation Meter - SRM 3006. The measurement experimental scenario followed the guidelines of the recommendation ΕΛΟΤ EN 61566, IEC 61566 [26-02-1999]. Based on this analysis, the main technical characteristics of the transmitter output, receiver input and antennae electrical & electromagnetic characteristics were estimated.

In order to estimate the average level of local mean strength at the receiver, we employed two different path loss models [4] - [6]. Antenna heights are considered to be sufficient to provide LOS and avoid blockage of local foliage. The terrain irregularities need to be accounted for, however, and these losses will be incorporated in a zero-mean Gaussian variable (in dB) for the large-scale variations of the local mean value of the received

power. For our calculations, the receiving antenna gain was not taken into consideration since we are interested in isolating the propagation phenomena from any possible variation of gains depending on whether an omni-directional antenna of low gain (2 dBi) or a high-gain directional antenna (9 dBi) will be employed [7] - [12].

We calculated as a sum of two independent processes: the distance-dependent path loss which is a deterministic loss due to free space propagation (provided by the idealistic Friis equation) [9], and the 'excess path loss', defined by Jakes as "the difference (in decibels) between the computed value of the received signal strength in free space and the actual measured value of the local mean received signal" [10].

The Free Space Model accounts only for the distance-dependent losses whereas the Log-Distance model incorporates the excess path loss as well, which, in this scenario, since LOS is considered, stands for the losses due to terrain and other geographical irregularities.

The average path loss (in dB) is provided by the following formula [13] for the Free Space Model:

$$P_L = 32.45 + 20\log_{10} f(MHz) + 20\log_{10} d(km) \quad (1)$$

The mathematical expression of the Log-Distance path loss model is given by [3]:

$$L_{total} = PL(d_0) + N\log_{10}\left(\frac{d}{d_0}\right) + X_\sigma \quad (2)$$

Where $PL(d_0)$ is the path loss at the reference distance, usually taken as (theoretical) free-space loss at 100 m, for outdoor propagation scenarios, N is the path loss distance exponent and Xs is a Gaussian random variable with zero mean and standard deviation of σ dB. N and σ are derived from experimental data. During our work a coverage probability of 95% was assumed and thus:

$$X_\sigma = z \times \sigma(dB) = 1.645 \times \sigma(dB) \quad (3)$$

Both models are suitable [1] for open areas unlike other models which are more practical for urban areas, such as the Hata and Okumura model [14] - [15].

In the case examined in this work, the path loss exponent assumes a value of 2 so as to express the free-space distance-dependent attenuation phenomena without incorporating any other losses. The zero-mean Gaussian variable is employed to express the 'excess path loss' and is set, in this paper, to a value of 6 dB so as to reflect losses due to terrain irregularities.

The distance required for signal propagation (T-R separation is approximately 38 km) calls for the employment of specific wireless technologies such as WiMax or LTE. Since in Greece LTE is provided in band 3 (1800 MHz), our calculations will focus on that frequency band. A bandwidth of 20 MHz is assumed and a medium-range LTE Base Station (BS) is assumed, with a maximum transmit power of 38 dBm. Since the LTE receiver sensitivity level for band 3 is -91 dBm, our main interest is to investigate the boundaries of reliable signal reception for distances close to, and beyond, the estimated T-R separation. Free Space model is more idealistic whereas the Log-Distance model may be more pessimistic but provides a "worst-case assumption" which is important when planning such long-distance links.

In our calculations, the antenna gain of the receiver unit was not included. The reason is that we want to provide estimations of the local mean value of the received signal at the "close proximity" of the receiver so as to provide numerical estimations without becoming dependent on the gain and directivity of receiver antennae. Results are depicted in Fig. 2 for both models.

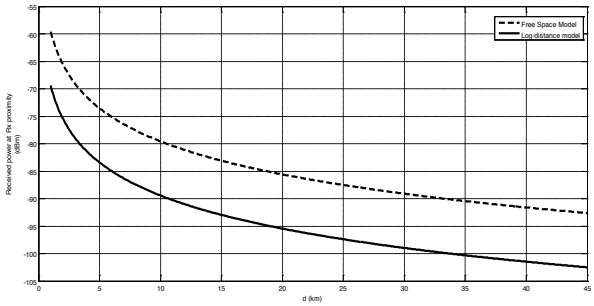

Fig. 2. Local mean value of received power as a function of distance

Figure 2 confirms that a dominant LOS component is essential for signal reception near the threshold. In the case of LOS propagation, where clearance of th 1st Fresnel zone is guaranteed, signal reception above receiver sensitivity level can be accomplished, thus allowing for a significant signal enhancement if a directional antenna is employed at the receiving end of the link. It should be reminded that in Fig. 2 results do not include antenna gain since the detection of signal levels above the sensitivity level in the proximity of the receiving antenna is investigated. Log-normal excess path loss can be discarded in a LOS-dominant propagation topology where the Free Space Model can provide reliable prediction of the local mean value of the received power (as a function of distance).

If scattering losses or any other excess path loss that does not comply with the distance-dependent free space attenuation assumption need to be considered, then signal reception drops below sensitivity level for LTE (-91 dBm) and outage occurs. In this case, the BS needs to be wide-area instead of medium-range.

In the case of low SNR region as the one examined in our case study, missing data can be a common occurrence. The development of a robust mathematical method for mitigating the missing values problem is of great significance in establishing a reliable system of wireless telemetry.

4. Imputation of missing data

In this study we propose a novel approach to the problem of missing values. Instead of applying the conventional linear regression models a new method which selects the best for prediction fitting model is followed [16], [17], [18] and [19]. The derived models used for imputing the missing values are therefore models optimized for prediction. They can best predict on the average the value which lies on a certain interval with some probability. This is achieved with the help of a mathematical quantity named beta expected tolerance regions. While the ordinary regression works with an extrapolation or interpolation of the best model fitting the data, the proposed technique lies on a probabilistic concept and suggests that model which best predicts a missing value within experimental region.

There are three distinct intervals in statistical analysis of data. For the problem of determining a missing value fitting a parameter to a model, the accuracy or precision may be described as a confidence interval, a prediction interval or a tolerance interval which are quite distinct.

Confidence intervals suggest how well the best-fit parameter determined by regression has been estimated. The crucial property is that confidence intervals provide information regarding the likely location of the true population parameter. On the other hand prediction intervals are different. They inform where next value sampled is expected to be found. The important point is that the prediction interval concerns the distribution of values, not the uncertainty in determining the population parameter.

The more demanding interval is the tolerance interval. It is defined on the basis of two percentages. The first percentage expresses "how sure" it is required the value to be and the second one represents what fraction of the values the interval will contain.

In our case a general linear model (GLM) can be expressed as $Y = T\beta + \sigma\varepsilon$ where $Y \in M_{n\times 1}$ is the observed vector of responses, $T \in M_{n\times(m+1)}$ is a matrix of known constant time intervals based on the m input variables and $\beta \in M_{(m+1)\times 1}$ with β being a vector of unknown parameters. Moreover $\varepsilon \in M_{n\times 1}$ is an unobserved random vector of errors and $\sigma^2 > 0$ unknown. Note that $M_{a\times b}$ is the set of $a \times b$ matrices. The unobserved random vector of errors ε satisfies $E(\varepsilon) = 0$, $E(\varepsilon\varepsilon') = I_n$ with $0 \in M_{n\times 1}$ a vector of zeros and $(I_n)_{ij} = \delta_{ij}$ is the unit matrix. In the contexts under discussion it is a safe assumption that errors follow a normal distribution with mean zero and variance 1. The coefficients β that have been modeled as a vector are determined by the least squared method. This is allowed due to the Gauss – Markov theorem. The least squared estimators will be determined from

$$\hat{\beta} = (T'T)^{-1}T'Y \tag{4}$$

where from now on, Y denotes a given realization of Y. The proposed algorithm is based on two strong and rich in content theorems that follow. The first theorem [20] allows us to evaluate the β-expectation structural tolerance region. Instead of using the concept of tolerance region it is better to use its improvement, the so called β-expectation structural tolerance region. For central 100β% of normal distributions being sampled it is given by the interval

$$\left[T'_{o,p}\hat{\beta}' - C(p)(S^{-1}_{1,p})^{-1/2}, T'_{o,p}\hat{\beta} + C(p)(S^{-1}_{1,p})^{-1/2}\right] \tag{5}$$

with

$$T'_{o,p} = (1, t, t^2, ..., t^p) \tag{6}$$

$$S^{-1}_{1,p} = I - T'_{o,p}(T'T + T_{o,p}{}'T_{o,p})^{-1}T_{o,p} \tag{7}$$

$$C(p) = t_{n-p}(\beta/2)(n-p)^{-1/2}(RSS_p)^{1/2} \tag{8}$$

Here, t_{n-p} denotes the student distribution with n-p degrees of freedom, and finally

$$\mathrm{RSS} = \left(\mathrm{Y} - \mathrm{T}\,\hat{\beta}\right)'\left(\mathrm{Y} - \mathrm{T}\,\hat{\beta}\right) \tag{9}$$

The prediction distribution can be evaluated as well with the help of the same theorem.

A second important theorem [17] that is going to be used, provides for the linear model the β-expectation tolerance region *Q(T, Y)*,

$$Q(T,Y) = \{\mathrm{w} \in M_m : \left(\mathrm{w} \simeq T^*\,\hat{\beta}\right)'S(T)\left(\mathrm{w} \simeq T^*\,\hat{\beta}\right) \le \frac{m}{n-p}F_{m,n-p1,\tilde{\beta}}\} \tag{10}$$

and

$$S(T) = \mathrm{I_m} + T^*(T'T)^{-1}T^{*\prime} \tag{11}$$

with $F_{\mathrm{m,n-p1},\tilde{\beta}}$ the $\tilde{\beta}$ quantile of the F distribution with m and n-p degrees of freedom. Note that the matrix T* corresponds to the matrix of the input observations of the missing values.

The novel algorithm for the missing value problem we are a going to present is written in pseudocode. It is a modification of the algorithm described in [19]. It consists of five steps.

The first step is to define matrix Y. Y in our case comprises PH, Radon, Redox, temperature and conductivity. All missing values Y* have to be determined, as well as their corresponding time values. In the presented code in Appendix A, we place each of the five factors that are components of Y in one dimensional matrices. If there are some missing values we have just to omit them and continue consecutively keeping the time ordering. The values of time at which measurements were taken comprise matrix T. They have to be placed in a list too. The time at which a measurement was lost, is omitted. In addition, the code

normalizes data T in the interval [-1,1] and the transformed data are placed in the list named "T'TRN".

The second step comprises the evaluation of the vector θ with the estimators of GLM model of *p-th* order polynomial. For every *p=0* to *k* we should evaluate from Eq. (4), estimators $\hat{\beta}$.

The third step is about the determination of the largest volume of the β-expectation tolerance region since this is the worst case of the input variables set, as far as prediction concerns. Then, in the fourth step we find the minimum β-expectation tolerance region among the worst models. This model with min β-expectation tolerance region is also the best one.

Therefore, in the third step we have for all p=0 up to k, to define matrices $T_{o,p}$ from Eq. (6). In the presented realization of the algorithm we have set k=10. Furthermore, it is essential to evaluate the length L_p at the point t_o. Following (5), (10) and working for the one variable degree polynomial the length of the tolerance region can be derived equal to

$$\mathrm{L_p(t_o)} = 2\mathrm{t_{n\text{-}p;1-\beta/2}}\left(\mathrm{n-p}\right)^{-1/2}\mathrm{RSS}^{1/2} \left\{\left(\mathrm{I}-\mathrm{T'_{op}}\left(\mathrm{T'T}+\mathrm{T'_{op}}T_{\mathrm{op}}\right)^{-1}T_{\mathrm{op}}\right)^{-1}\right\}^{1/2} \quad (12)$$

with the point t_0 being the point that gives maximum value as follows

$$\mathrm{Max}\left[\mathrm{T'_{o,p}}\left(\mathrm{T'T}\right)^{-1}\mathrm{T_{o,p}}\right] \text{ with } -1 \le t \le 1 \quad (13)$$

The pseudocode names as "LTR" the length of the tolerance region Eq. (12) while expression named "quantity" is the one that has to be maximized at time t_o.

The fourth step is realized with a loop that runs from p=0 to as much as 10 in the function named "ORDER". This function chooses the minimum of the stored maximum "lengths" and finds the corresponding *p* value which is the degree of the response function for the best predictive model. It also compares the best predictive model with the ordinary best fitting model finally, evaluating $\mathrm{RMS} = \mathrm{RSS}/\sqrt{\mathrm{n-p-1}}$. The best fit model according to the conventional method is the one with the minimum RMS.

The fifth and final step refers to the evaluation of all the missing values based on the best predictive model. This becomes possible with the help of the code calling the function named "MISSINGVALUES".

For all measured quantities the proposed method was applied for evaluating the missing values according to the best for prediction polynomial. This was done for several different time durations. Results reveal that for many datasets the best fitting for prediction polynomial is different from the one suggested from ordinary methods with distance criteria like RMS, and therefore, the imputation of missing values is different as well. If the designer of an experiment is interested in correct predictions then the discussed method gives distinctive and correct results.

The applied algorithm differs nontrivially from all other methods that impute missing data with a distance criterion. Furthermore, it was tested that it suggests in many cases different values than those derived with the ordinary methods. The numerical study of this methodology proved that the algorithm, which chooses a polynomial model according to the best prediction criterion, does not have the disadvantage to select the largest order polynomial, as the best model. This is a well known problem associated with methods using distance criteria. Finally, data analysis with respect to the proposed method showed that for data with large dispersion, the estimated missing data are different from the values suggested by an RMS criterion.

5. Discussion

For the LOS wireless system a study was performed in order to find out the link budget parameterization. Local mean values of the received signal at the "close proximity" of the receiver (without considering any specific antenna gain for the receiver unit) is found to be approximately equal to the receiver sensitivity level of -91 dBm, if free space (distance-dependent) losses are considered. If terrain and/or foliage irregularities are taken into considered in a generic assumption of a 6 dB shadow depth, then the local mean value of the received signal suffers an additional attenuation of approximately 10 dB, resulting in a value below -100 dBm. Since we have already assumed a worst-case path loss of 150 dB [21], all possible attenuation scenarios have been accounted for.

Site-specific measurements and a more thorough investigation of plantation, foliage, and their effect on signal propagation with regard to variable antenna heights will validate the robustness of each assumption. Furthermore, the use of a high-gain directional antenna at the receiver unit will provide the necessary signal enhancement (in the order of 9 dB) so that the signal will be well above the sensitivity levels. Finally, it should be noted that the maximum transmit power of 38 dBm concerns a medium-range LTE BS that applies for micro-cell scenarios. LTE specifications explicitly provide a macro-cell (suburban-open areas) option of a wide-area BS where no upper bound in maximum transmit power was considered (and will be subsequently provided by local and/or regional band regulations by respective authorities). Therefore, our calculations simply provide scenarios that range from sub-optimal to worst-case in order to test the robustness of signal propagation under specific conditions, leaving room for improvement on both transmitter and receiver side, as well as addressing issues for future work in the channel characteristics, as heralded by other published works [22] - [24].

6. Results

The testing of the proposed algorithm with the measured data reveals an affirmative conclusion for using the proposed method. There is a strong theoretical background [16], [17], [18] and [19] that ensures the success of the method to any applied field. It was found that for several datasets the proposed novel method for missing data imputation differs non trivially from the other existing methods. In the present study a three cycles evaluation scheme was selected for the adjustment of the whole measuring station including the sensors, the control unit as well the transmission and the data recorder. The first and second evaluation periods lasted 25 days each. The third evaluation period lasted 30 days. These three time periods were consecutive started on March 2014. During the first period we had 2.3%, 2.5% , 3.4%, 0.5% and 2.8% of missing values for PH, Radon, Redox, temperature and conductivity respectively. In this first period there was a two days out of operation time duration that was excluded. In the second period we had 2.1% , 2.9%, 2.3%, 0.3% and 3.8% of missing values for PH, Radon, Redox, temperature and conductivity respectively while in the third period we got 2%, 1.9%, 2.4%, 0.3% and 1.9% missing values. Utilizing the developed algorithm we managed to impute the missing values in all the fifteen time series referring to the measured five factors of these three evaluation periods. The derived best for prediction polynomials each for every one of the five factors were then used to model the time evolution and calibrate the measuring station for best performance. The reliability of the measurement of each factor and the reliability of the whole measurement process were evaluated. The reliability (intraclass correlation coefficient, average measures at 95% CL) of the third period of evaluation reached 0.83, 0.77, 0.84, 0.95 and 0.75 for PH, Radon, Redox, temperature and conductivity respectively. In addition, using the derived best fitting polynomials we were able to calculate various statistical inferences regarding trends, and correlations which will be part of a larger statistical analysis of the measurements.

For research programs under the scope to utilize data for making safe and scientifically concrete predictions, it is not considered reasonable and appropriate to build models based on curve fitting as closely as possible to the data. Instead, it is desirable for the model to be determined by a curve that fits data with the help of a best fit polynomial for the prediction of the Y value for a certain X (within the experimental region), establishing a specific degree of high probability.

It would also be interesting as a future research to generalize the proposed method for problems with multiple independent variables or for cases like [25] and [26]. Another interesting investigation is to develop an algorithm capable to handle both time series analysis and missing data imputation.

Appendix A

The present section, presents the pseudo-code that realizes the algorithm.

(*Here we set data for matrix T. The time intervals that correspond to the missing data should be omitted, for example {1,2,3,4,6,7,8} for a missing value at t=5*)

1. Set an one dimensional matrix TIME`INTERVALS with elements consecutive integers that represent the time intervals at which measurements were performed
   (*Here we set all data for matrix Y which in our example concerns radon, PH, conductivity or OPR. *)
2. Set an one dimensional matrixY that consists of all time ordered measurements of factor Y omitting the missing values
3. Define the function/routine that estimates the t student distribution probability density function for the relevant tolerance region
   tst[n_] := Sqrt[-n + (1/n*(0.05*(Sqrt[n]*Beta[n/2, 1/2]))^(2/(1 + n)))^(-1)];
   (*Here we transform adequately the data*)
4. Set to variable T`TRN a normalized value in the interval (-1,1) i.e.
   T`TRN := (TIME`INTERVALS - A)/B;
5. Set variable “n” as the number of all time intervals
6. Calculate in variable A and B the quantities A := 1/2 (UU + DD);
   B := UU - A;
7. Estimate the minimum and maximum values of time intervals and set them in variables DD and UU respectively
   DD = Min[TIME`INTERVALS];UU = Max[TIME`INTERVALS];
8. Set a 2 X N “TRNdata” matrix with first column the transformed time intervals and second column the relevant Y factor. N is the number of all time intervals.
   (* the main code*)
9. Define the coefficients T[0],T[1],...T[10] of the ten polynomials which will be tested with both criteria
   T[0] := {1}
   T[1] := {1, T`TRN
   T[2] := {1, T`TRN, T`TRN^2}
   ...
   Set each of the 10 polynomials of order “i” Top[i]=1+t+t^2+...t^i
10. Set the quantity that should be maximized, see Eq.(13)
    quantity := $Top^T. (T^T.T)^{-1}.Top$;
11. Define a subroutine function “Maximize” which finds the value of t inside the normalized region[-1,1] that maximizes distance “quantity”
12. Define the useful expression EXPR, see equation (7), which is used in the definition of the “probabilistic length” Lp
    EXPR := $[1 - Top^T. [T^T.T + Top.Top^T]^{-1}. Top]^{-1}$;
13. Define the useful quantity “bi” that is used in the definition of the criterion RMS
    bi := $[T^T. T]^{-1}.(T^T.Y)$;

14. Define the quantity RSS (see equation (9)) used in the definition of the criterion RMS
    RSSp := [Y - T. bi]^{T} . (Y - T. bi);
15. Define the quantity"LTR" which is the length of the tolerance region and it is evaluated for the t that maximizes "quantity". This is the prediction criterion
    LTR := 2*tst[n- i] *(n - i - 1)^(-1/2)
    ((EXPR)^(1/2))(RSSp)^(1/2);
16. Define the expression that gives the conventional RMS criterion for best fitting model
    RMS := RSSp/(n - i - 1);
17. Define a subroutine/function returning a report order by order which provides justification about the best fitting polynomial. The user must set as an argument for this function the largest order of the polynomial to be tested. The maximum possible order that can be selected is 10. The prediction criterion LP and the conventional criterion RMS are estimated for each order of the polynomial. In addition the function returns for each order of the polynomial the plot of the data together with the best fitting polynomial for prediction.
    START LOOP
    For i=0 to n Maximize[quantity, -1 <= t <= 1}, t]];
    g1 = ListPlot[TRNdata];
    g2 = Plot[Top^{T}.bi, in the interval {t, -1, 1}];
    Print[Top^{T}.bi];
    Find the value t and name it "maxquant" that maximizes "quantity"
    maxquant := NMaximize[quantity, [respecting the constraint -1 <= t <= 1}] ;
    Print["RMS=", RMS];
    Set t = maxquant;
    Print["LP=", LTR];
    Print[Show[g1, g2, PlotRange -> All]]
    END LOOP
18. Define the subroutine/function that returns the missing value, the order of the best fitting polynomial used for estimating the missing value and the minimum tolerance length. The user sets as first argument the time that corresponds to the missing value and the largest order of the polynomial to be tested
    Set the time of missing value in the variable "timeofevent"
    Calculate tt = (timeofevent - A)/B;
    Set the maximum order of the polynomials to "nn"
    START LOOP
    For i=0 to nn
    Find the value t and name it "maxquant" that maximizes "quantity"
    maxquant := Maximize[quantity, {respecting constraint -1 <= t <= 1}] ;
    t = maxquant; LTRmin := LTRR[0]
    order := i ;
    IF LTRR[i] <= LTRmin; THEN LTRmin := LTRR[i]
    t = tt;
    MISSINGVALUE = Top[order]^{T}.bi;
    Print["MISSING VALUE=", N[MISSINGVALUE],
    " order of best polynomial=", order, " minimum LP=", LTRmin]
    END LOOP

Acknowledgment

Authors acknowledge that this work has been financially supported by the research program, entitled: "Measurement of Environment Physical-Chemical Parameters by Development Autonomous Data Collection Processing Transmission Systems with use of green Power and most optimal management", MIS 380360, within the research activity "Archimedes III", funded by the NSRF 2007-2013.

References

[1] GLASGOW, H. B, BURKHOLDER J. M., REED R. E., LEWITUS, A. J, KLEINMAN, J. E.: *Real-time remote monitoring of water quality: a review of current applications, and advancements in sensor, telemetry, and computing technologies*, Journal of Experimental Marine Biology and Ecology, Volume 300, Issues 1–2, pp. 409-448, 31 March 2004.
[2] MADDALA, G.: Introduction to Econometrics, 2nd Edn. Macmillan : New York, pp.: 663,1992
[3] STIGLER, S. M.: *Gauss and the Invention of Least Squares*. The Annals of Statistics, vol. 9, No. 3, 465-474, 1981.
[4] GOLDSMITH, A.: Wireless Communications. Cambridge : Cambridge University Press, 2005.
[5] PARSONS, J. D.: The Mobile Radio Propagation Channel. Hoboken : NJ : Wiley Interscience, 2000.
[6] RAPPAPORT, T.: Wireless Communications: Principles & Practice. *Upper Saddle River*, NJ: Prentice Hall, 1999.
[7] CHRYSANTHOU, C. , BERTONI, H. L.: Variability of Sector Averaged Signals for UHF Propagation in Cities, *IEEE Transactions on Vehicular Technology*, vol. 39, No. 4, pp. 352-358, November 1990.
[8] ERCEG, V. , GREENSTEIN, L. J. , TJANDRA, S. Y. , PARKOFF, S. R. , GUPTA, A. , KULIC, B. , JULIUS, A. A., BIANCHI, R.: An Empirically Based Path Loss Model for Wireless Channels in Suburban Environments, IEEE *J. on Selected Areas in Communications*, vol. 17, No. 7, July 1999.

[8] ODA, Y., TSUCHIHASHI, R., TSUNEKAWA, K., HATA, M.: *Measured Path Loss and Multipath Propagation Characteristics in UHF and Microwave Frequency Bands for Urban Mobile Communications,* Vehicular Technology Conference, 2001. VTC 2001 Spring. IEEE VTS 53rd, vol. 1, 6-9 May 2001 pp. 337-341 vol.1.

[10] SALO, J., VUOKKO, L., EL-SALLABI, H. M., VAINIKAINEN, P.: An Additive Model as a Physical Basis for Shadow Dading, *IEEE Transactions on Vehicular Technology*, vol. 56, No. 1, pp. 13-26, January 2007.

[11] CHRYSIKOS T., KOTSOPOULOS, S.: *Site-specific Validation of the Path Loss Models and Large-scale Fading Characterization of Large-scale Fading for a Complex Urban Propagation Topology at 2.4 GHz*. The 2013 IAENG Intern. Conference on Communication Systems and Applications (IMECS 2013), March 13-15, 2013, Hong Kong.

[12] SEYBOLD, J.: Introduction to RF Propagation. Hoboken, NJ : Wiley Interscience, 2005.

[13] JAKES, W. C. (Ed.): Microwave Mobile Communications. New York : Wiley Interscience, 1974.

[14] HATA, M. : Empirical Formula for Propagation Loss in Land Mobile Radio Services, IEEE Transactions on Vehicular Technology, vol. 29, No. 3, pp. 317-325, August 1980.

[15] OKUMURA, Y., OHMORI, E., KAWANO, T., FUKUDA, K. : Field Strength and its Variability in VHF and UHF Land-Mobile Radio Service, *Review of the Electrical Communication Laboratory,* vol. 16, No. 9-10, pp. 825-873, September-October 1968.

[16] KITSOS, C. P.: An Algorithm for Construct the Best Predictive Model. Softstat' 93: Advances in Statistical Software, Faulbaum, F. (Eds.). Stuttgart : New York, pp. 535-539, 1994.

[17] ELLERTON, R. R. W., KITSOS, C. P., RINCO, S.: Choosing the Optimal Order of a Response Polynomical-structural Approach with Minimax Criterion. *Commun. Stat. Theory Meth.*, 15: 129-136, 1986.

[18] MULLER, C. H., KITSOS, C. P.: Optimal Design Criteria Based on Tolerance Regions. mODa 7-Advances in Model-Oriented Design and Analysis, Bucchianico, A., H. Lauter and H. P. Wynn (Eds.). Physica-Verlag, pp. 107-115, 2004.

[19] Kitsos, C. P., V. Zarikas, "On the Best Predictive General Linear Model for Data Analysis: A Tolerance Region Algorithm for Prediction", Journal of Applied Sciences 01/2013; 13(4):513-524. DOI:10.3923/jas.2013.513.524, 2012

[20] HAQ, M. S., RINCO, S.: β-Expectation Tolerance Rfor a Generalized Multivariate model with Normal Error Variables, *J. of Multivariate Analysis*, 6 (3), pp. 414-421, 1976

[21] ZARIKAS, V., CHRYSIKOS, T., ANAGNOSTOU, K., KOTSOPOULOS, S., AVLAKIOTIS, P., LIOLIOS, G.T. LATSOS, C. PERATZAKIS, LYGDIS, A., LYKOYRGIOTIS, A. A.: *Telemetry, Analysis and Wireless Data Communications for a Measuring Station*, Elektro 2014 10th Intern. Conference, May 2014, Rajecké Teplice

[22] MICEK, J., KARPIS, O.: Wireless Sensor Networks for Road Traffic Monitoring. *Communications - Scientific Letters of the University of Zilina*, vol. 12, pp. 80-85, 2010, ISSN 1335-4205.

[23] HODON, M., PUCHYOVA, J., KOCHLAN, M.: Smartphone-Based Body Area Network for Stress Monitoring: Radio Interference Investigation. *Communications - Scientific letters of the University of Zilina*, ISSN 1335-4205.

[24] BRIDA, P., MATULA, M., DUHA, J.: Using Proximity Technology for Localization in Wireless Sensor Networks. *Communications - Scientific Letters of the University of Zilina*, ISSN 1335-4205, vol. 9, No. 4, pp. 50-54, 2007.

[25] GIKAS, V., STRATAKOS, J.: A Novel Geodetic Emethod for Accurate and Automated Road/railway Centerline Geometry Extraction Based on the Bearing Diagram and Fractal Behavior, *IEEE Trans. Intell. Transp. Syst*., 13: 115-126, 2012.

[26] ZARIKAS, V., GIKAS, V., KITSOS, C. P.: Evaluation of the Optimal Design "Cosinor Model" for Enhancing the Pof Robotic Theodolite Kinematic Observations, *Measurement*, vol. 43, No. 10, December 2010, pp. 1416-1424.