Stochastic model and optimum sampling interval of variables for environmental measurement systems

Obiora Sam Ezeora
Electrical Engineering and Informatics
University of Pardubice
Pardubice, Czech Republic
obiora.sam.ezeora@student.upce.cz

Jana Heckenbergerova
Inst. of Mathematics and Quantitative Method
University of Pardubice
Pardubice, Czech Republic
Jana.heckenbergerova@upce.cz

Petr Musilek
Electrical and Computer Engineering
University of Alberta
Edmonton, Canada
petr.musilek@ualberta.ca

***Abstract*— It is common in engineering to model time-dependent variables as diffusion process represented by stochastic differential equations. This is usually helpful when empirical datasets describing time evolution of variables are available. This helps in accurate estimation of parameters of the stochastic differential equation which describes the dynamic system. Additionally, it helps in characterization and determination of optimal performance of the system. The above have been conducted in this study using real environmental field data. Linear stochastic model was fitted to longitudinal datasets and optimum sampling interval investigated. A new method has been proposed for determination of optimum sampling interval. Results obtained differ from those of hypothetical optimum which do not take energy consumption into consideration.**

Keywords*— *Stochastic; autoregressive; longitudinal data; time series; energy consumption; sampling interval; environmental variable

I. INTRODUCTION

The application of stochastic differential equations (SDE) and stochastic partial differential equations (SPDE) in modeling and analysis of electronic and embedded systems is gaining huge attention. In wireless communication systems, SDE models are developed and used to analyze networks [1]. SDE models are also been used to model and analyze electrical power networks and systems [2, 3, 4 and 5].

For measurement systems, limited work involving stochastic differential analyses exist. This study is timely now there is strong focus on energy-efficient ways of determining optimum sampling interval of measurement systems. While it helps to realize an energy-efficient system, it also contributes in the integration and use of electric power from low-grade energy sources.

In complex systems which operate with numerous measurement systems, such as autonomous vehicles, determining optimum sampling interval and filtering of noise emanating from sensors are vital. They help in the development of algorithms required for effective and efficient operations of the system. Determining optimum sampling interval and noise filtering complement each other. Optimum sampling interval requires that accurate model and its statistical estimates be computed. They represent important statistical properties of the variable, in the mist of noisy and partial sensor observations. Consequently, they require modern stochastic filtering techniques. Studies relating to dynamic stochastic processes cannot be separated from this. It is on this framework that this study is based.

In this study, a stochastic method of modeling longitudinal datasets is discussed. It also proposed a new method of determining energy-efficient optimum sampling interval. Impacts of various sampling intervals on model estimates were also discussed. Hypothetical optimum sampling interval is also discussed.

At this juncture, it shall be noted that sampling interval as used in this context refers to sampling time-interval. It is the time interval separating successive samplings or observations.

Finally, this study consists of eight sections. Section I provides background information relating to problem statement, objectives and areas of applicability of the results. Section II provides the mathematical background. Section III provides information on sampling design and characterization of data used in the analysis. The methodology is discussed in section IV while results obtained are explained in section V. Determination of energy-efficient (optimum) sampling interval is discussed in section VI while impacts of using different sampling intervals are discussed in section VII. Concluding statements are presented in section VIII.

II. MATHEMATICAL PRELIMINARIES

A stochastic differential equation (SDE) can be represented as shown in equation (1).

$$dX_t(\omega) = f_t\big(X_t(\omega)\big)dt + \sigma_t\big(X_t(\omega)\big)dW_t(\omega) \text{ [6]} \qquad (1)$$

Deterministic part of the SDE which is represented by function, f, is called drift. The function σ_t, represents diffusion coefficient while $dW_t(\omega)$ represents the noise term.

Equation (1) is a diffusion process and follows a continuous path. Therefore, equation (1) can be better represented with an

This work has been mainly supported by the University of Pardubice via SGS Project (reg. no. SG660026) and institutional support, and partly supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

978-1-5090-6406-9/17/$31.00 ©2017 IEEE

integral, such as Ito integral. This will help discretize the continuous path for solutions. The implementation involves numerically methods such as Euler-Maruyama or Milstein approximation schemes so that SDE coefficients are estimated.

However, in order to reduce complexity and iteration run-time without sacrificing on accuracy, both drift and diffusion terms in equation (1) were considered as first order polynomials. This reduces equation (1) to linear stochastic differential equation shown in equation (2).

$$dX_t(\omega) = \big(\alpha_t + \beta_t X_t(\omega)\big)dt + \vartheta_t dW_t(\omega)\ [6, 7] \qquad (2)$$

Where α, β, ϑ represent deterministic functions.

Equation (2) can be written in a clearer form of linear stochastic differential equation. This is shown in equation (3) [6, 7].

$$X_{t+1} - aX_t = b + \rho Z_{t+1}, \quad t = 0, 1, 2, \ldots \qquad (3)$$

Where:
a, b, ρ are scalars and always greater than zero.

X_0 is represented by x_0, and $\{Z_t\}$ is a set of exogenous stochastic process [6, 7, 8].

$\{Z_t\}$ consists of independent and normally distributed residuals with mean of zero and finite variance.

Linear stochastic differential equation shown in equation (3) above is also known as an autoregression process of first order. That is, equation (3) is an AR(1) equation [6, 7, and 8].

It then follows that equation (3) is a special type of AR(1) process which requires errors to be independent and identically distributed (iid). If errors are not iid, they are modeled so as to fulfill the iid requirement. This is a mandatory requirement for linear regression models of which equation (3) also represents.

At this juncture, it should be noted that errors and residuals have been used interchangeably in this context. They refer to difference between measured value and model-predicted value. Errors as used in this context do not refer to measurement error (i.e. difference between measured value and true value).

III. Sampling Design And Methodology

Field data used in this study were obtained through measurements taken within a residential area in the city of Edmonton, Alberta. Three sensors were deployed at three locations within the site for measurement of photosynthetically active radiation (PAR). The sensor nodes were installed in such a way that certain important environmental conditions were represented. One PAR sensor was installed near a fence so that partial shading of the fence cast over it.

Another sensor was installed under a canopy formed by couple of trees. The third sensor was installed so that it receives solar radiation directly with no shadow or canopy over it. Another sensor node was used for air temperature measurements. Data were collected from May 3, 2015 to July 2, 2015 with intervals between measurements changing from 5 seconds to 30 seconds in several steps (5, 10, 15, and 30). The sensor nodes were powered by a solar harvester supplemented by a supercapacitor and backup primary batteries.

In this longitudinal data study, all datasets between 09:35:15 and 22:52:45 were considered daytime datasets while those between 22:53:15 and 09:34:45 were considered nighttime datasets. This reduces data inhomogeneity caused by commercial and industrial noise during diurnal cycle. As a result, more accurate results were obtained.

IV. Methodology

The methodology consists of following steps:

1. Use field data to perform ordinary regression so that estimates of coefficients and their standard errors could be determined. Durbin Watson statistical estimate is also computed.
2. Analyze Durbin Watson statistical estimate to determine if residuals are uncorrelated (independent).
3. Validate results from Durbin Watson method with those obtained using residuals chi-square independent test, residuals autocorrelation function (ACF) and partial autocorrelation function (PACF) analyses.
4. If residuals are autocorrelated (i.e. dependent), the assumption of linear regression modeling is violated. The variables are then modeled using Prais-Winsten method. It iterates until the estimates converge. This removes autocorrelated errors and presents adjusted estimates of the regression model together with their corresponding standard errors. Due to page limitations, the mathematics explaining re-estimation and adjustment of model could not be presented.
5. Prais-Winsten method was considered over Exact Maximum-Likelihood and Cochrane-Orcutt due to its simplicity and lower computational time. In addition, Prais-Winsten method can eliminate the usual effects due to absence of data for first lag residual.
6. Obtained model is validated by comparing predicted values with measured values. Additionally, model-fit performance parameters such as coefficient of determination (R^2), mean-square-error (MSE) were analyzed.

Above steps were repeated using longitudinal datasets from different sampling interval. This was done so as to determine the impacts of varying sampling interval on standard error of estimates.

V. Results And Discussions

Table I shows statistical estimates of daytime and nighttime air temperature models with sampling interval of 30 seconds.

From Table I, it would be seen that Durbin Watson estimate for both daytime and nighttime air temperature is zero. This indicates that daytime air temperature values are positively autocorrelated. The same situation applies to nighttime air temperature values. This is further supported by residuals autocorrelation function (ACF) and partial autocorrelation function (PACF) plots shown in figures 1-4.

TABLE I. STATISTICAL ESTIMATES OF DAYTIME AND NIGHTTIME AIR TEMPERATURE MODELS WITH SAMPLING INTERVAL OF 30S

Variable	Coefficients		Standard error of coefficients	Durbin Watson
Daytime air temperature (ºC)	Intercept	26.813	0.451	0.000
	Slope/ Predictor	0.004	0.000	
Nighttime air temperature (ºC)	Intercept	11.392	0.166	0.001
	Slope/ Predictor	-0.001	0.000	

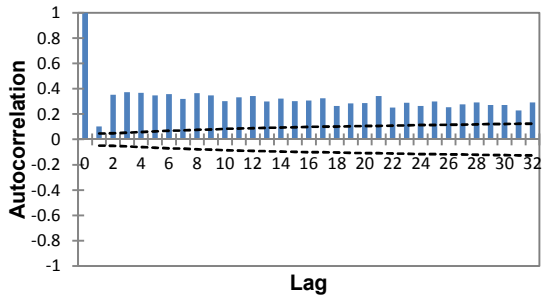

Fig. 1. Residuals ACF of daytime air temperature

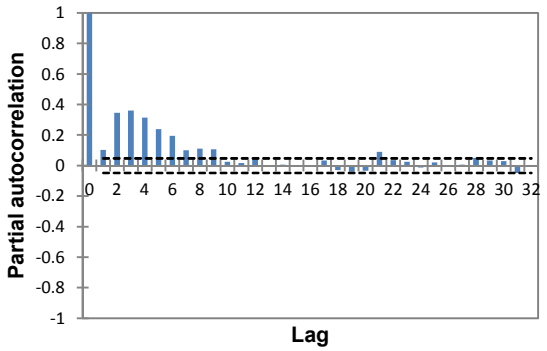

Fig. 2. Residuals PACF of daytime air temperature

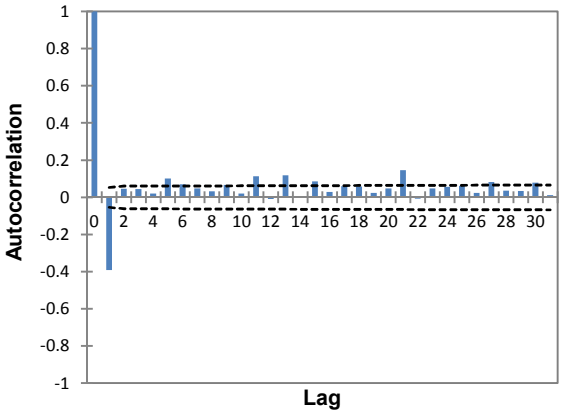

Fig. 3. Residuals ACF of nighttime air temperature

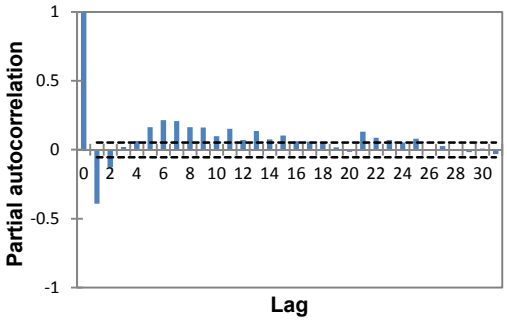

Fig. 4. Residuals PACF of nighttime air temperature

Residuals ACF and PACF plots show that one partial autocorrelation is statistically significant. This implies the residuals are autocorrelated. It also shows that the residuals have an AR(1) structure. Therefore, it violates the assumption that residuals should be independent.

Table II shows adjusted statistical estimates of daytime and nighttime air temperature regression models when sampling time-interval is 30 seconds. It should be noted that all estimates used in this study (context) are unstandardized.

TABLE II ADJUSTED STATISTICAL ESTIMATES OF DAYTIME AND NIGHTTIME AIR TEMPERATURE MODELS WITH SAMPLING INTERVAL OF 30S

Variable	Estimates		Standard error of estimates	Durbin Watson
Daytime air temperature (ºC)	Intercept	17.154	16.804	1.797
	Slope/ Predictor	-0.002	0.002	
	Rho (AR(1))	1.000	0.001	
Nighttime air temperature (ºC)	Intercept	16.250	6.468	2.785
	Slope/ Predictor	0.001	0.001	
	Rho (AR(1))	1.000	0.001	

From Table II, it would be seen that Durbin-Watson values are 1.797 and 2.785 respectively. When looked up at Durbin-Watson significance table, they were found to be within the corresponding lower and upper bounds. This indicates that the positive autocorrelated errors have been removed. This is further supported by ACF and PACF plots of the residuals. Those for nighttime air temperature are shown in figures 5 and 6 respectively.

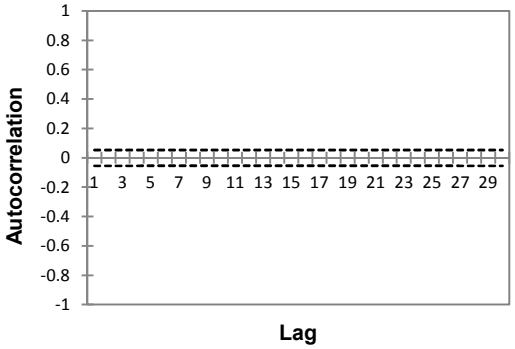

Fig. 5. Residuals ACF plots for adjusted nighttime air temperatures

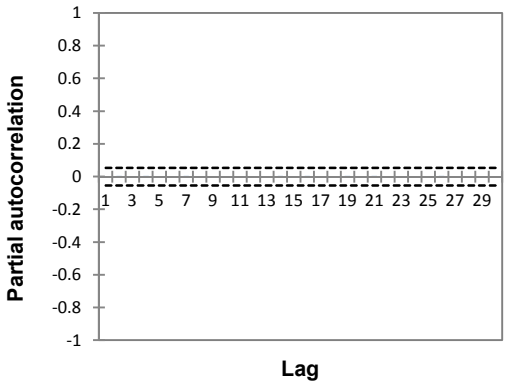

Fig. 6. Residuals PACF plots for adjusted nighttime air temperatures

It would be seen from figures 5 and 6 that no significant autocorrelation or partial autocorrelation exists at any of the lags. Considering the generalized equation shown in equation (3), model equations were deduced based on estimated coefficients shown in Table II. The model equations are shown in equations (4) and (5) respectively.

$$D_t = 17.154 - 0.002D_{t-1} + 1.000Z_t \quad (4)$$

$$N_t = 16.250 - 0.001N_{t-1} + 1.000Z_t \quad (5)$$

D_t represents air temperature measured at daytime instant, t.

D_{t-1} represents air temperature measured at daytime instant, t-1. Such as air temperature value observed 30 seconds earlier during the daytime (if sampling interval is 30 seconds).

N_t represents air temperature measured at nighttime instant, t.

N_{t-1} represents air temperature measured at nighttime instant, t-1.

Z_t represents independent and normally distributed residuals with mean of zero and finite variance.

VI. Determination Of Optimum Sampling Interval

In order to determine optimum sampling interval, concept of tolerance- uncertainty relationship was adopted. Tolerance is well known as the total acceptable uncertainty. It is also known as the maximum permissible error (MPE) of measured value. It is noted by international standards that tolerance should not be more than three times the uncertainty of instrument measuring the variable [9, 10, 11]. Uncertainty of the instrument measuring the variable is available in the calibration certificate of the instrument. Consequently, the maximum permissible error or tolerance for measured value can be estimated.

It then follows that all values of a variable that are within tolerance range are acceptable based on international standard. Therefore, optimum sampling interval is that maximum sampling time-length within which values are still within the acceptable tolerance range of previously measured value. For example, considering data on technical datasheet for Hobo sensor S-THB-M002, its maximum permissible error should be (± 0.2)*(3) = ± 0.6 ºC. Hence, measured temperature value of say, 20ºC has temperature values within 20 ± 0.6ºC acceptable as accurate and close enough to nominal value of 20ºC.

The question becomes: "why should sampling interval of temperature involving this instrument be set so that several sampling within this range (20 ± 0.6ºC) are performed. Air temperature values that lie within the range (20 ± 0.6ºC) are considered good enough. They can represent the nominal value (20ºC). By so doing, contributes in securing energy-efficiency.

However, in systems where sensitivity to small temperature changes is mandatory input for microprocessor decisions, several temperature measurements within the tolerance range are usually required. For such systems, instrument with lower uncertainties and energy consumption should be considered. Alternatively, temperature values could be predicted from measured values. This obviously requires a model. Model development method and steps discussed in this study can be used. Interpolation technique for non-integer time indexes is also required. This has been discussed in [12].

Optimum sampling interval is therefore estimated by finding a local maximum time index that can allow new measured values to fall within the acceptable tolerance of previously measured value. Table III shows time series excerpt in comparison with its model-predicted values. Model methodology described in this study has been used in predicting the values. Table III is also used to demonstrate how proposed optimum sampling interval would be determined for an example involving Hobo sensor S-THB-M002.

TABLE III. Optimum Sampling Interval – Daytime Air Temperature

Date/Time	Measured value (ºC)	Predicted value (ºC)	Residuals (ºC)	Remark
2015-06-01 09:35:45	17.0[A]	17.0601	-0.0601	Tolerance: 17.0 ± 0.6 ºC
2015-06-01 09:36:15	17.0	17.0600	-0.0600	Still within tolerance of A
2015-06-01 09:36:45	17.0	17.0600	-0.0600	Still within tolerance of A
2015-06-01 09:37:15	17.1	17.0600	0.0400	Still within tolerance of A
Continues till next row	Continues till next row	Continues till next row	Continues till next row	Still within tolerance of A
2015-06-01 09:48:15	17.6	17.5984	0.0016	Still within tolerance of A
2015-06-01 09:48:45	17.7	17.5984	0.1016	Measured value out of (17±0.6ºC) range. New sampling required.

From Table III, it would be seen that in most cases, air temperature values only changed by 0.1ºC in 30 seconds or more. This applies to all daytime air temperature measurements taken during the entire measurement period. For daytime measured value of 17.0 ºC shown in Table III, sampled values obtained after 12 minutes 30 seconds were still within the acceptable tolerance of 17.0 ºC. This is validated by repeating above analysis using other measured values within the data

series. These repeated analyses also helped in ensuring that measured value used in determination of optimum sampling interval represents true value. For nighttime air temperature, optimum sampling interval was longer. For the system described above, 390 seconds is an energy-efficient (optimum) sampling interval. For systems where small temperature changes are required, variables may be sampled at intervals below the energy-efficient sampling interval.

VII. Hypothetical Optimum Sampling Interval

Table IV and fig. 7 show impacts of varying sampling interval on standard error of intercept and that of autocorrelation coefficient (rho). The same time span (length) was used in the analysis.

TABLE IV. Effects of varying sampling interval on model estimates of daytime air temperature

Sampling interval (seconds)	Standard error of slope (°C)	Standard error of intercept (°C)	Standard error of autocorrelation coefficient (rho), °C
5	0.000	0.001	1.613
10	0.001	0.000	22.128
15	0.001	0.000	21.880
30	0.002	0.001	16.000

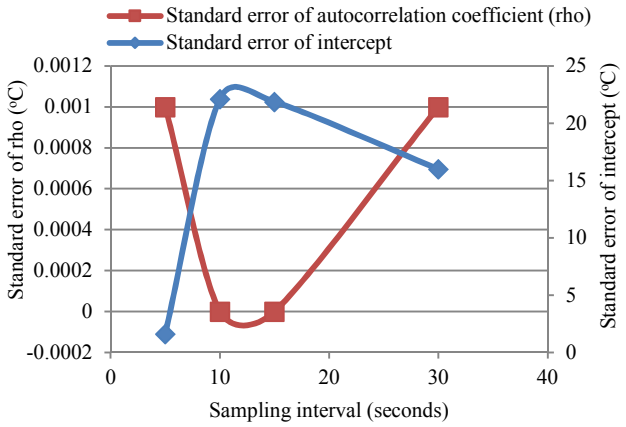

Fig. 7. Sampling interval effects on model estimates of daytime air temperature

It would be seen that standard error of intercept increases as sampling interval increases until it reaches its maximum value. At this point of maximum standard error of intercept, standard error of autocorrelation coefficient is at its minimum. Sampling interval corresponding to point of minimum standard error of autocorrelation is the hypothetical optimum sampling interval. It is hypothetical because it is based on mathematical theories and conditions. It does not consider external factors such as energy consumption of the measuring device. It only considered the mathematical properties of datasets. The hypothetical optimum sampling interval is not energy-efficient. It is for this reason that the analysis presented in preceding section was performed.

Similar study was performed by Åstrom [13]. In [13], it was shown that variance of the drift parameter (deterministic component) of a time series SDE model decreases with increasing sampling interval. It continues to decrease until a local minimum, known as hypothetical optimum sampling interval is reached. Further increase in sampling interval beyond the local minimum results in rapid increase in variance of the drift term.

VIII. Conclusions

In this study, simple and accurate method for modeling longitudinal data of a time series has been discussed. The method iterates for convergence with less run-time. New method of determining energy-efficient (optimum) sampling interval was also discussed. The method requires that few measurements be taken. This enables repeat of analysis using different measured values.

The study also investigated impacts of varying sampling interval on standard error of estimates. It was found that increasing sampling interval decreases standard error of autocorrelation coefficient until minimum value is reached. Thereafter, standard error increases with increasing sampling interval. This minimum sampling interval is described as hypothetical optimum sampling interval because it considers only mathematical properties of datasets. The proposed new method is therefore recommended.

References

[1] J. Shea, L. Zacharious, and B. Pasik-Duncan, "Computational methods for stochastic differential equations and stochastic partial differential equations involving standard Brownian and fractional Brownian motion," Challenges of Modern Technology, vol. 2, issue 2, 2011.

[2] H. Verdejo, et al, "Stochastic modeling to represent wind power generation and demand in electric power system based on real data," Journal of Applied Energy, vol. 173, no. 1, pp. 283-285, July 2016.

[3] E. Kolarova and L. Brancik, "Vector stochastic differential equations used to electrical networks with random parameters," International Journal of Advances in Telecommunications Electrotechnics Signals and Systems, vol. 2, no. 1, 2013.

[4] K. M. R. Kumar and P. K. Chenniappan, "Stochastic differential equation modeling for electrical networks," Journal of Computer Applications, vol. 2, no. 2, April – June, 2009.

[5] T. K Rawat and H. Parthasarathy, "Modeling of an RC circuit using a stochastic differential equation", Thammasat International Journal of Science, vol. 13, no. 2, April – June 2008.

[6] J. Wang and S. Wang, Business Intelligence in Economic Forecasting: Technologies and Techniques, Information Science Reference, Hershey, Pennsylvania, USA, 2010.

[7] C. Edmond, "Advanced Macroeconomic Techniques" , Lectcure Notes, assessed online at: http://pages.stern.nyu.edu/~cedmond/406/N4A.PDF

[8] Z. Horvath and R. Johnson, AR(1) Time Series Process, Econometrics 7590, assessed online at: http://www.math.utah.edu/~zhorvath/ar1.pdf

[9] ISO 10012-1, Quality Assurance Requirements for Measuring Equipment - Part 1, International Organization for Standardization, Geneva, Switzerland, 1998. ISBN 0 580 41654 2.

[10] ISO 22514-7, Statistical Methods in Process Management – Capability and Performance, Part 7: Capability of Measurement Process, Geneva, Switzerland, 2012.

[11] OIML R 111, International Organization for Legal Metrology, Paris, France, 2004.

[12] O. Ezeora, J. Heckenbergerova and P. Musilek, A new adaptive sampling method for energy-efficient measurement of environmental parameters, 16th IEEE International Conference on Environment and Electrical Engineering, Florence, Italy, (2016) June 7-10.

[13] K. J. Åstrom, "On the choice of sampling rates in parametric identification of time series," Journal of Information Sciences, vol 1, issue 3, July 1969, pp. 273-278.