The Seventh Asian-Pacific Symposium on Structural Reliability and Its Applications (APSSRA2020)
October 4–7 2020, Tokyo, Japan
T. Takada, I. Yoshida & T. Itoi (editors)

Regressions with Dummy Variable to Tackle with Environmental and Operational Influences on Long-term Bridge Health Monitoring

W.J. Jiang[1], C.W. Kim[1,*], Y. Goi[1] and F.L. Zhang[1,2]

[1] *Department of Civil & Earth Resources Engineering, Graduate School of Engineering, Kyoto University, Japan.*
[2] *School of Civil and Environmental Engineering, Harbin Institute of Technology, Shenzhen, China.*
[*] *E-mail: kim.chulwoo.5u@kyoto-u.ac.jp*

Abstract: In long-term structural health monitoring (SHM), the environmental and operational variables (EOVs) such as temperature, vehicle and wind may introduce interference into the identified damage indicators (basically like modal frequencies, strain, displacement, etc.), and sometimes even cause a masking effect to damage impact. This kind of EOVs-related variability decreased the sensitivity and reliability of damage indicators. This study is intended to model the EOVs-related variability in the identified damage indicators utilizing the Gaussian process regression (GPR) technique combined with dummy variable. Considering the unclear nonlinear and coupling mechanism of EOVs-related impacts, GPR was applied and compared with some classical linear regression methods. A dummy variable was introduced to improve the performance of regression on the lack of observations. Investigations on the performance of different regression methods using residuals showed that GPR presented a better performance in capturing the EOVs-related variability, and a proper dummy variable as a supplementary resulted in an improvement in GPR.

Keywords: deficient measurements, dummy variable, EOVs-related variability, Gaussian process, long-term SHM.

1. Introduction

The long-term structural health monitoring (SHM) based on system identification and pattern recognition techniques has played a more and more important role in the assessment and maintenance of infrastructures. However, how to deal with uncertainties have been a major challenge in the long-term SHM. As one of the major sources of uncertainty, environmental and operational variables (EOVs) such as temperature, humidity, vehicle and may induce apparent variation (sometimes make the damage impact blurry) in damage indicators such as modal properties. This kind of uncertainties could decrease reliability of decision based on SHM.

Cornwell et al. (1999) investigated the environmental variability of modal properties with the data from the Alamosa Canyon Bridge, and found there were significant changes in modal frequencies along with the variation of temperature during a 24-hour period. Zhou and Sun (2019a, 2019b) investigated the environmental and operational effects in a sea-crossing bridge and a cable-stayed bridge, respectively, and clarified the underlying mechanism of the related impacts to modal frequencies, displacement, tower distance, etc. Comanducci et al. (2015) utilized an analytical parametric model of suspension bridge to study the impact of wind loading on modal frequencies compared with damage effect. The result showed that frequency variations caused by changing wind speed can be more significant than those produced by a small damage.

To address the impacts of EOVs in SHM, Nandan and Singh (2014a, 2014b) introduced a state space model-based approach for the correlation research between modal frequency and temperature, and proposed two kinds of filtering methods to remove seasonal trend in observations. Ubertini et al. (2017) utilized a multiple linear regressive filter to remove temperature effects from identified modal frequencies and carried out an assessment of structural health condition based on novelty detection in the residuals. In recent years, various methods based on Bayes' theorem have been studied to model the EOVs-related variability. Kim et al. (2018) proposed a Bayesian approach considering multiple factors such as temperature and vehicle weights in a long-term SHM on a Gerber-type steel girder bridge. Mu and Yuen (2018) developed an updated version of the sparse Bayesian learning for the regression of the relation between modal frequency and environmental condition. Avendaño-Valencia and Chatzi (2020) combined a Gaussian process with vector autoregressive model in different time-scales to model the variation of structural dynamics under varying wind speed and ambient temperature.

As a basic means for modeling EOVs-related variability in long-term SHM, the regression analysis is usually adopted to recognize the correlative pattern and make a prediction. However, due to the deficient measurements, unclear coupling effect and linear approximation, to build a model with good performance in both reproduction and generalization is generally intractable. Therefore, this study investigates the performance of Gaussian process regression (GPR) compared to classical linear regression methods under the condition of deficient measurements and unclear coupling effect. In addition, a way to introduce dummy variable as a supplementary for deficient measurements was discussed associated with a case study as well.

2. Regression methods

2.1 Classical linear regression

Classical regression models generally adopt a scheme of the linear regression with a basic form shown in Eq. 1.

$$y = \alpha X + \varepsilon \tag{1}$$

where y denotes the response variable and $X=\left[1,x_1,x_2,\cdots,x_n\right]^T$ denotes the predictor variables or explanatory variables; $\alpha=\left[\alpha_0,\alpha_1,\alpha_2,\cdots,\alpha_n\right]$ is the coefficient vector to be estimated, and ε denotes the residual error which is usually assumed to be independent and identically distributed (IID).

For the ordinary least square regression (OLSR), the objective function equals to the loss function as follows.

$$J(\alpha)=\frac{1}{n}\left\|y-\alpha X\right\|_2^2 \tag{2}$$

where $\|\cdot\|_2$ denotes the L2 norm. Then, coefficient estimator $\hat{\alpha}$ is calculated by minimizing $J(\alpha)$ as follows.

$$\hat{\alpha}=\underset{\alpha}{\arg\min}\ J\left(\alpha\right) \tag{3}$$

In order to control the complexity of the regression model, regularization skills are often utilized by adding a penalty term. Least square shrinkage and selection operator (LASSO) is one of these methods frequently utilized in the linear regression analysis. The objective function is given by Eq. 4.

$$J(\alpha)=\frac{1}{n}\left\|y-\alpha X\right\|_2^2+\lambda\left\|\alpha\right\|_1 \tag{4}$$

where $\|\cdot\|_1$ denotes the L1 norm and λ is a shrinkage factor (or called as regularization parameter) which can be decided with a cross-validation.

Compared with other regularization skills like the ridge regression (or called Tikhonov regularization), LASSO has an additional effect in sparse feature selection and can help further decrease the complexity of the regression model.

2.2 Gaussian process regression (GPR)

The GPR is a nonparametric and fully Bayesian regression method (Rasmussen and Williams 2006). GPR describes the regression model as Eq. 5 treating *f*(*X*) as the latent function without a fixed form, and estimates the distribution of *f*(*X*) based on the Bayesian inference.

$$y=f(X)+\varepsilon \tag{5}$$

Here, *f*(*X*) is supposed to be a stochastic process with a priori as follows.

$$f(X)\sim GP[0,K(X,X)] \tag{6}$$

where $K(X,X)=\mathrm{E}\left[f(X)f(X)\right]$ is a kernel covariance matrix; $X=\{X_1,X_2,\cdots X_m\}$ and $y=\{y_1,y_2,\cdots,y_m\}$ are the sample sets for model training with a set size *m*. If $\varepsilon\sim N(0,\sigma_n^2)$ is the noise with IID, and *K*(.) denotes a kernel covariance matrix, the predictive posterior distribution has a solution as follows.

$$p(f_*\left|X,y,X_*,\sigma_n^2)=N[f_*\right|\bar{f}_*,\mathrm{cov}(f_*)] \tag{7}$$

$$\bar{f}_*=K(X_*,X)\left[K(X,X)+\sigma_n^2 I\right]^{-1}y \tag{8}$$

$$\mathrm{cov}(f_*)=K(X_*,X_*)-K(X_*,X)\left[K(X,X)+\sigma_n^2 I\right]^{-1}K(X,X_*) \tag{9}$$

where *X** denotes the inputs of test samples, and *f** is the corresponding predicted outputs.

To get the posterior distribution, hyper-parameters including signal variance, length scale and error variance should be determined by means of the maximum likelihood estimation (MLE) and other techniques like cross-validation. Compared to classical regression methods, an advantage of the GPR is that the model is not confined to be linear and the solution space contains all possible forms of nonlinearity if enough kernel functions are considered. In addition, the assumption of the GPR priori is not required to be IID which is generally a restriction in the statistical linear regression.

2.3 Dummy variable

A dummy variable is the one that has only the value 0 or 1 to indicate the existence of categorical effect that may make a difference in the output of a model. It is also called a qualitative variable and mainly used in the linear regression analysis. For a dataset that can be divided into *m* groups by a qualitative attribute, a basic form of the linear regression with the dummy variable can be described as Eq. 10.

$$y=\alpha X+\beta D+\varepsilon \tag{10}$$

where $D=\left[D_1,D_2,\cdots,D_{m-1}\right]$ is a vector of dummy variables; $\beta=\left[\beta_1,\beta_2,\cdots,\beta_{m-1}\right]$ is the corresponding coefficient vector, and the rest components are same as Eq. 1. For a dataset that has just two different group values in qualitative attribute, only one dummy variable will be introduced into the model.

As deficient measurement of EOVs is a universal problem in SHM, involving dummy variable based on the acquired observations may help improve the performance of regression under this condition. Since the GPR differs from the linear regression method, a trial of combining the GPR with dummy variable is also considered in this study.

2.4 Jensen-Shannon divergence

The Jensen-Shannon divergence offers a way for measuring the similarity between two probability distributions in statistics, which is also known as information radius in the information theory. It is an improvement of the Kullback-Leibler (KL) divergence and takes a basic form as follows.

$$D_{JS}(P\,\|\,Q)=\frac{1}{2}D_{KL}(P\,\|\,M)+\frac{1}{2}D_{KL}(Q\,\|\,M) \tag{11}$$

where *P* and *Q* are the compared probability distributions, and $M=\frac{1}{2}(P+Q)$.

The LHS of Eq. 11 is the notation of Jensen-Shannon divergence, while the RHS is a combination of two KL divergences with a basic form for discrete probability distribution as follows.

$$D_{KL}\left(P\,\|\,M\right)=\sum_{x\in\chi}P\left(x\right)\log\left(\frac{P\left(x\right)}{M\left(x\right)}\right) \tag{12}$$

where *P* and *M* are defined in the same probability space χ.

Since the residuals of training set and validating set are usually investigated in regression analysis to assess the performance of a regressed model, the Jensen-Shannon divergence thus can be utilized to assess the generalization ability of a model by measuring the similarity between distributions of residuals in training set and validating set.

3. Case study

As a case study, data from a long-term monitoring on a simply supported steel plate-girder bridge with reinforced concrete deck are investigated. Fig. 1 shows the monitoring bridge with a span of 40.5 m and width of 4.5 m, which was built in 1957. The monitoring on the bridge was started from September 1st, 2016, and the operation of this bridge was ceased from March, 2017 since the operation of a newly built bridge nearby was started.

3.1 Data pre-processing

The bridge was mainly monitored with a sensor system of accelerometers and thermometers. The sampling frequency of accelerometers was set as 200 Hz, while the temperature was recorded every 30 minutes. Details of the sensor setup can be seen in Kim et al. (2019).

Based on the data collection, a Bayesian FFT method was utilized to identify the first three bending modal frequencies per 30 minutes (see Fig. 2 for results). The average modal frequency of each mode was 3.11 Hz, 9.50 Hz and 21.91 Hz, respectively, and an obvious trend was observed in the time series of the first and second bending modes.

Temperature is the only measured EOVs (see Fig. 3), and with a feature selection process (stepwise regression and LASSO), the temperature records ‘T_2’, ‘T_4’ and ‘T_5’ were reserved to be explanatory variables in the regression analysis.

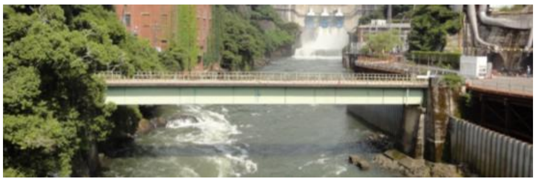

Figure 1. Elevation view of the target bridge.

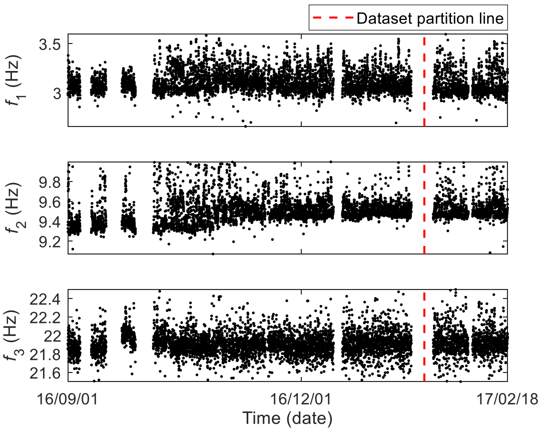

Figure 2. Identified modal frequencies (f_1 to f_3: 1st to 3rd bending frequencies; partitioned to training and validating datasets).

3.2 Regression analysis and comparison

Based on above data collections, a regression analysis is carried out with OLSR, LASSO and GPR, respectively. An Engle-Granger two-step method for co-integration test is firstly conducted with an OLSR model, and the feasibility of classical regression methods is suggested. The temperature-related varying patterns are then trained and validated on separate datasets. For LASSO, a cross-validation process is applied to select optimal shrinkage factor. For GPR, the hyper-parameters including kernel parameters and variance of noise are estimated by MLE and optimal values were subsequently selected from the estimated solution space by the cross-validation.

Figure 4 shows the predictive results on validating set based on the mentioned three methods. Compared with OLSR and LASSO, the results indicated that GPR captured more local features of the variation in the dataset. A difference in the predictability among three bending modal frequencies is observed from the plot. Compared to the 1st and 2nd modal frequencies, prediction of the 3rd modal frequencies seemed to be intractable. Two explanations may dedicate to this situation: (i) the 3rd modal vibration is hard to be excited by ambient excitations with a higher modal frequency around 21.91 Hz, which means the identified 3rd modal frequencies may contain more noisy components; (ii) the higher modes may be less sensitive to temperature. Thus, regression analysis about the 3rd modal frequency was ignored in further investigation.

To further assess the performance of each regressed model, an investigation about the properties of residuals were carried out while the residuals of the 3rd modal frequency were ignored (see Fig. 5). As is widely acknowledged in regression analysis, the ability of reproduction and generalization is declared for a good regressed model. Therefore, some indices like mean square error ‘*MSE*’, coefficient of determinant ‘R^2’ and normalized Akaike Information Criterion ‘*nAIC*’ were calculated for assessing the performance of reproduction (see Table 1).

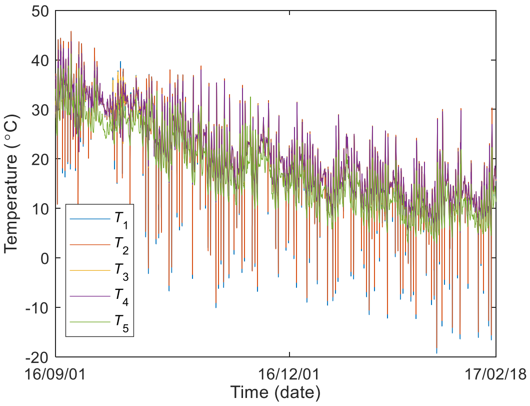

Figure 3. Temperature records in five sensors (T_1 to T_4 were attached to two girders, and T_5 was outside the bridge).

The results in the table also proposed GPR showed a better goodness of fit of the training set. As to the ability of generalization, an optimal model should leave the residual as a white process, or at least a stationary process. However, as shown in Fig. 5, an apparent distinction of the residual distribution between training set and validating set was observed in three models, which indicated a non-stationary property. This phenomenon may be caused by the model approximation, intermittent measurement and deficient EOVs involved, with the last one considered as a crucial factor. According to Chang et al. (2014), existence of vehicles might cause increment or decrement in identified modal frequencies of bridges, especially in low order modes of short-span bridges like the one in this research.

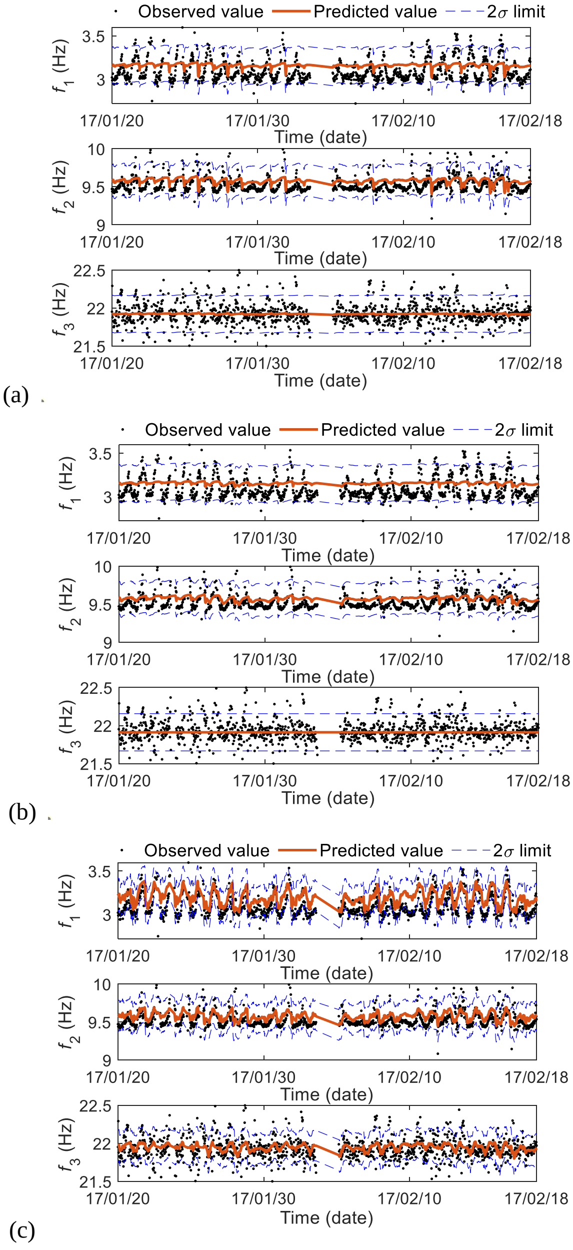

Figure 4. Prediction on validating set: (a) OLSR; (b) LASSO; (c) GPR.

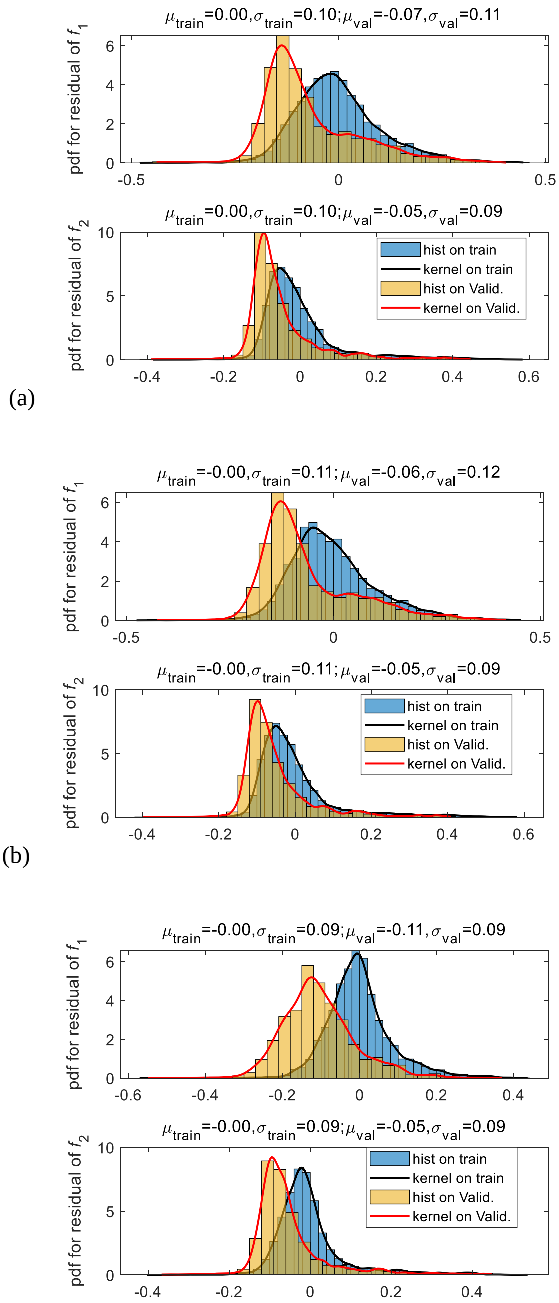

Figure 5. Comparison of residual distribution between training set and validating set: (a) OLSR; (b) LASSO; (c) GPR.

Table 1. Indices for assessment of reproduction ability.

Response	Method	MSE	R^2	nAIC
1st modal frequency	OLSR	0.0108	0.1181	-4.5253
	LASSO	0.0110	0.1007	-4.5057
	GPR	0.0074	0.3950	-4.9021
2nd modal frequency	OLSR	0.0109	0.2936	-4.5150
	LASSO	0.0113	0.2718	-4.4846
	GPR	0.0081	0.4775	-4.8165

A superposition of thermal effect and traffic effect exactly dedicated to the phenomenon of trend in the time series of the 1st and 2nd modal frequencies as previously mentioned. Thus, an idea about introducing dummy variable to reflect the existence of vehicle at each observation was investigated in next discussion.

3.3 Combination with dummy variable

Since traffic effect may be an influential factor affecting the variability of modal frequencies and traffic condition was not recorded in this case study, the existence of vehicle at each observation may be a qualitative attribute that makes a difference in the output of the model. Although it is hard to judge the existence of vehicle without relative records, diurnal period that vehicle may exist with a high probability may be roughly detected from these two reasons: (i) generally in common sense, the existence of vehicle in the daytime and nighttime must have different probability; (ii) the inconsistency of the variation extent in temperature and modal frequency may indicate a high probability of vehicle existence due to a superposition effect. Then, according to these two ideas, the diurnal period along with temperature increase is statistically investigated (see Fig. 6 (top plot)). Observations that presented an inconsistency of variation extent in the period corresponding to a temperature increasing process are roughly detected based on the differential of dataset and statistical quartiles, with corresponding periods shown in Fig. 6 (bottom plot).

From Fig. 6, it can be noted that the period from 9 a.m. to 5 p.m. may indicate a higher probability of vehicle existence than other clock time in a day. Therefore, a dummy variable is defined to reflect the difference of these two periods: (i) period from 9 a.m. to 5 p.m. with the value ‘1’ of dummy variable to corresponding observations; (ii) period outside this interval with the value ‘0’. Figure 7 shows a correlation plot between modal frequency and temperature. It can be found that this defined dummy variable grouped these observations with a good performance.

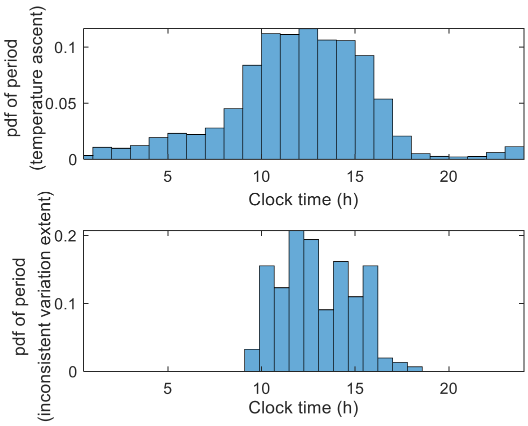

Figure 6. Histograms for statistical investigation on: period of temperature increase (top plot); period of inconsistent variation extent along with temperature increase (bottom plot).

The regression analysis with dummy variable is subsequently carried out. Comparison between the GPR and the GPR with dummy variable is shown in Fig. 8, and a better performance was observed in the prediction by the GPR with dummy variable. The residuals in the former three regression methods with dummy variable are also investigated (see Fig. 9), and apparent changes in the residual distribution can be noted. To compare the generalization ability before and after considering dummy variable in each method, the Jensen-Shannon divergence is calculated to reflect the distance between two distributions in training set and validating set. As shown in Table 2, the results proposed that introduction of the defined dummy variable reduced the distinction of the residual distribution between training set and validating set, and definitely improved the generalization ability of each method in this case.

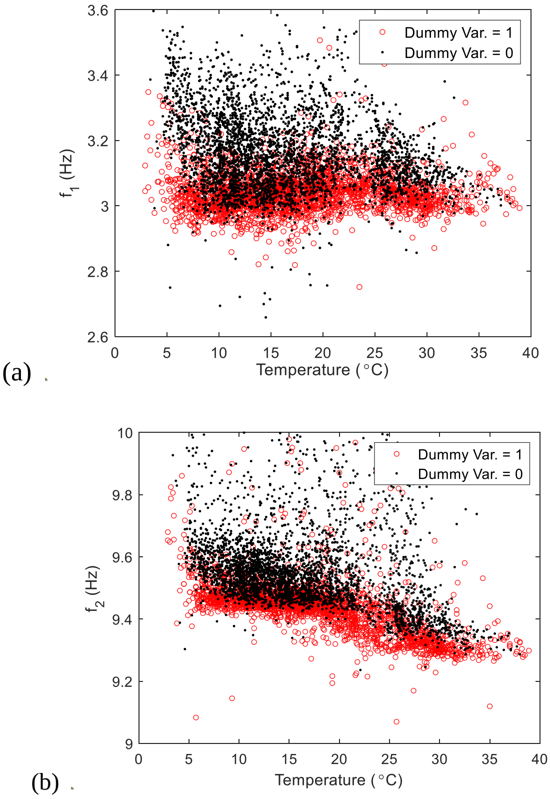

Figure 7. Correlation plot between modal frequency and temperature: (a) 1st modal frequency; (b) 2nd modal frequency.

Table 2. Jensen-Shannon divergence of the residual distribution between training set and validating set.

Response	**Method**	**Jensen-Shannon divergence**	
		WO/dummy	W/dummy
1st modal frequency	OLSR	0.1596	0.1301
	LASSO	0.1551	0.1057
	GPR	0.2871	0.2506
2nd modal frequency	OLSR	0.1500	0.1081
	LASSO	0.1092	0.0643
	GPR	0.1952	0.1547

4. Concluding remarks

Regression methods to reduce EOVs to long-term SHM are discussed. The dummy variable is introduced into the regression models to improve regression and prediction by reducing the influence of deficient measurements. Observations through this study can be summarized as follows.

(i) The data quality makes a different influence to classical linear regression methods and the GPR. Under the situation of deficient measurement, GPR captures more local features about the EOVs-related variability, while classical linear regression methods like OLSR and LASSO are awfully affected by the data quality.

(ii) Introducing proper dummy variable to further group the original dataset may improve the performance of regression methods in some extent, especially when the records of EOVs are deficient. However, to search and define a proper dummy variable needs further studies.

(iii) In addition to deficient measurement, model approximation and intermittent measurement may be also two factors that mainly affect the regression performance. The former one weighs more in classical linear regression as the reality of nonlinear and coupling effects, while the latter one may take more errors to GPR as it considers the covariance matrix of time series.

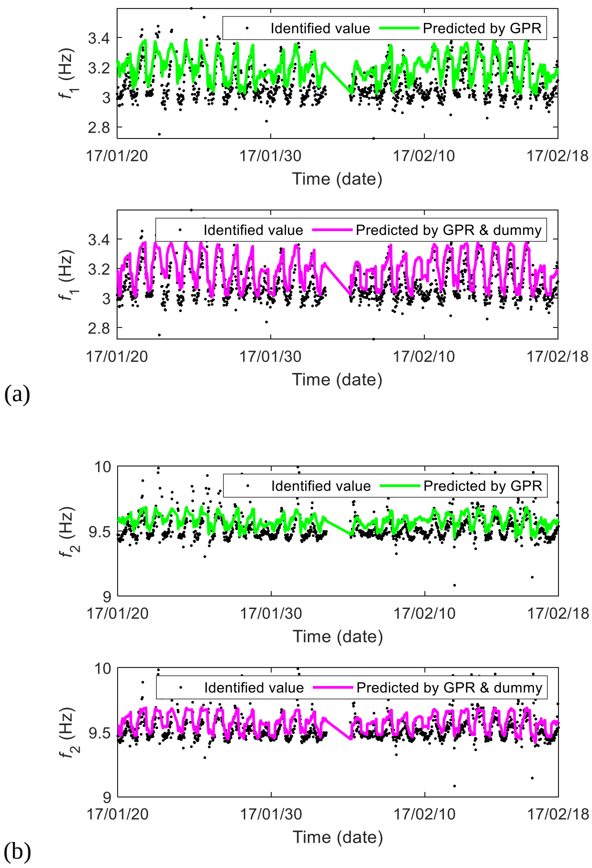

Figure 8. Comparison of GPR and GPR with dummy variable in prediction: (a) on 1st modal frequency; (b) on 2nd modal frequency.

Acknowledgments

This study is partly sponsored by a Japanese Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (B) under project No.19H02225. That financial support is gratefully acknowledged.

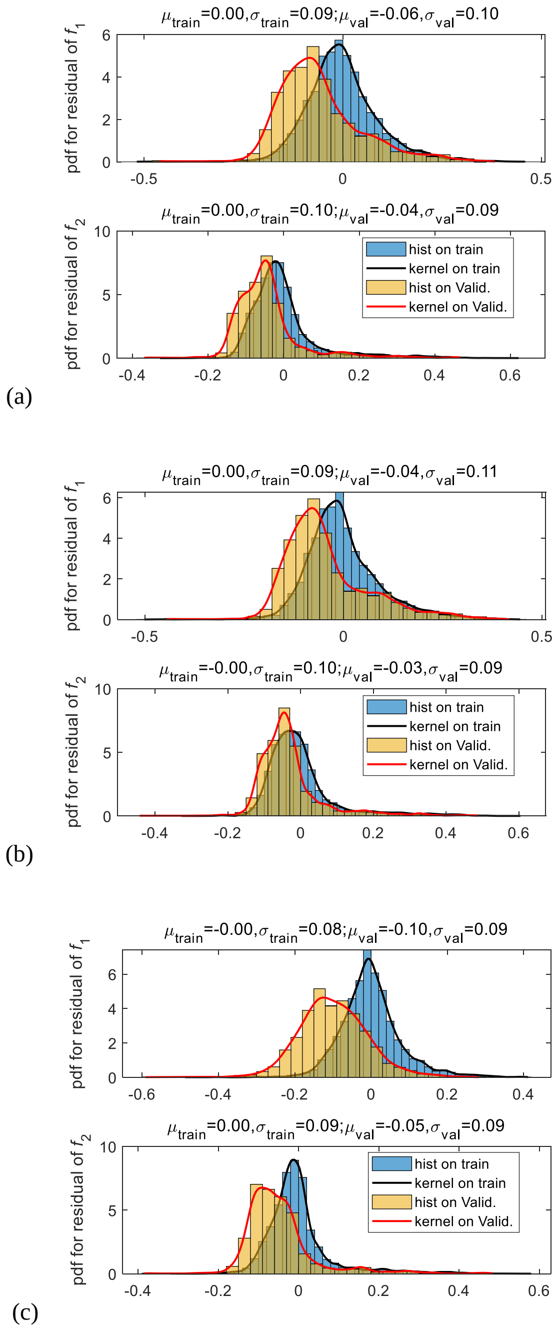

Figure 9. Comparison of residual distribution between training set and validating set involving defined dummy variable: (a) OLSR; (b) LASSO; (c) GPR.

References

Avendaño-Valencia, L.D. and Chatzi, E.N. 2020. Multivariate GP-VAR models for robust structural identification under operational variability. *Probabilistic Engineering Mechanics*, 60: 103035.

Cornwell, P., Farrar, C.R., Doebling, S.W. and Sohn, H. 1999. Environmental variability of modal properties. *Experimental Techniques*, 23(6): 45–48.

Comanducci, G., Ubertini, F. and Materazzi, A.L. 2015. Structural health monitoring of suspension bridges with features affected by changing wind speed. *J. Wind Eng. Ind. Aerodyn.*, 141: 12–26.

Chang, K.C., Kim, C.W. and Borjigin S. 2014. Variability in bridge frequency induced by a parked vehicle. *Smart Struct. Syst.*, 13(5): 755–773.

Kim, C.W., Zhang, Y., Wang, Z., Oshima, Y. and Morita, T. 2018. Long-term bridge health monitoring and performance assessment based on a Bayesian approach. *Struct. Infrastruct. Eng.*, 14(7): 883–894.

Kim, C.W., Hirooka, T., Goi, Y., Hayashi, G. and Mimasu, T. 2019. Influence of local damage and change in boundary condition on frequency changes of a steel plate girder bridge, *Proc. of PSSC2019*, Tokyo, Japan, Nov. 9-11, 2019.

Mu, H.Q. and Yuen, K.V. 2018. Modal frequency-environmental condition relation development using long-term structural health monitoring measurement: Uncertainty quantification, sparse feature selection and multivariate prediction. *Measurement*, 130: 384 –397.

Nandan, H. and Singh, M.P. 2014a. Effects of thermal environment on structural frequencies: Part I – A simulation study. *Engineering Structures*, 81: 480–490.

Nandan, H. and Singh, M.P. 2014b. Effects of thermal environment on structural frequencies: Part II – A system identification model. *Engineering Structures*, 81: 491–498.

Rasmussen, C.E. and Williams, C.K.I. 2006. Gaussian Processes for Machine Learning, MIT Press.

Ubertini, F., Comanducci, G., Cavalagli, N., Pisello, A.L., Materazzi, A.L. and Cotana, F. 2017. Environmental effects on natural frequencies of the San Pietro bell tower in Perugia, Italy, and their removal for structural performance assessment. *Mechanical Systems and Signal Processing*, 82: 307–322.

Zhou, Y. and Sun, L.M. 2019a. Effects of environmental and operational actions on the modal frequency variations of a sea-crossing bridge: A periodicity perspective. *Mechanical Systems and Signal Processing*, 131: 505–523.

Zhou, Y. and Sun, L.M. 2019b. A comprehensive study of the thermal response of a long-span cable-stayed bridge: From monitoring phenomena to underlying mechanisms. *Mechanical Systems and Signal Processing*, 124: 330–348.